Developing Deep Transfer and Machine Learning Models of Chest X-ray for Diagnosing COVID-19 Cases using Probabilistic Single-Valued Neutrosophic Hesitant Fuzzy

Hassan A. Alsattar[a,b], Sarah Qahtan[c], Aws Alaa Zaidan[d], Muhammet Deveci[e, f,*], Luis Martinez[g], Dragan Pamucar[h,*], Witold Pedrycz [i]

[a] *Department of Business Administration, College of Administrative Science, The University of Mashreq, 10021 Baghdad, Iraq. [email: hassan.alsattar@gmail.com]*
[b] *MEU Research Unit, Middle East University, Amman, Jordan*

[c] *Department of Computer Center, College of Health and Medical Techniques, Middle Technical University, 10047 Baghdad, Iraq (email: sarah.qahttan@gmail.com)*

[d] *SP Jain School of Global Management, Lidcombe, Sydney, NSW 2141, Australia (email: aws.alaa@gmail.com)*
[e] *Royal School of Mines, Imperial College London, London SW7 2AZ, UK (email: muhammetdeveci@gmail.com)*
[f] *Department of Industrial Engineering, Turkish Naval Academy, National Defence University, 34940 Tuzla, Istanbul, Turkey*

[g] *Department of Computer Sciences, University of Jaén, Jaén, 23071, Spain [email: martin@ujaen.es]*

[h] *Faculty of Organizational Sciences, University of Belgrade, 11000 Belgrade, Serbia (email: dpamucar@gmail.com)*

[i] *Department of Electrical and Computer Engineering, Faculty of Engineering, University of Alberta, 9211 116, Street NW, Edmonton, Alberta, T6G 1H9, Canada [email: wpedrycz@ualberta.ca]*

[*] Corresponding author

Developing Deep Transfer and Machine Learning Models of Chest X-ray for Diagnosing COVID-19 Cases using Probabilistic Single-Valued Neutrosophic Hesitant Fuzzy

Abstract

This study presents a novel dynamic localisation-based decision (DLBD) with fuzzy weighting with zero inconsistency (FWZIC) under a probabilistic single-valued neutrosophic hesitant fuzzy set (PSVNHFS) environment to benchmark Hybrid Multi Deep Transfer and Machine Learning (HMDTML) models. The novel DLBD method is proposed to generate a dynamic localisation decision matrix based on the upper and lower boundaries and the length of the scale. The superiority of DLBD derives from its ability to manage dynamic changes with boundary value consequences. In addition, the utilization of PSVNHFS in conjunction with DLBD and FWZIC has proven to effectively address the challenges posed by vagueness, uncertainty and hesitancy in the benchmarking procedure. The proposed methodology consists of three primary three steps: i) the adaptation of 48 HMDTML models, including 4 deep transfer learning models and 12 machine learning models trained on a dataset of 936 chest X-ray images obtained from both COVID-19 patients and individuals without the disease. Then, these models were evaluated based on seven evaluation criteria, and a decision matrix was proposed. ii) The development of a PSVNH–FWZIC to assign weights to the evaluation criteria. iii) The formulation of a PSVNH–DLBD for the purpose of benchmarking HMDTML models. Results of the PSVNH–FWZIC revealed that AUC and time were the most important evaluation criteria, while precision was the least important. Furthermore, the results from PSVNH–DLBD, reveal that Model M24 (Painters-Decision Tree) earned the highest rank when $\lambda = 2, 3, 4, 5$ and 6, followed by Model M25 (SqueezeNet-AdaBoost) and Model M34 (DeepLoc-kNN), while Model M39 (DeepLoc-SVM) had the lowest rank (*rank=48*) across all λ values. The proposed method underwent sensitivity and comparison analyses to confirm its reliability and robustness.

***Keywords*—** Chest X-Ray Images, COVID-19, Deep Transfer Machine Learning Models, FWZIC, and MCDM.

1 Introduction

Early diagnosis is an essential public health strategy in all situations because it allows for treatment at the earliest stage possible, increasing the likelihood of a positive outcome (Rong et al., 2020). Mortality rates can be dramatically lowered through detection and treatment at an early stage (Medhat & El Kassas, 2020). Although a conventional diagnosis provides a quick diagnostic process for patients, it may pose a significant risk to radiographers. Accordingly, medical imaging techniques such as X-ray and computed tomography (CT)-based screening are reasonably safe, efficient and accessible (Khammas et al., 2022;

Maghdid et al., 2021). CT and X-ray imaging methods have been widely used for a variety of disease screens, resulting in significant advancements in early diagnosis (Ammas et al., 2020; Ravi et al., 2021). However, X-ray scanning is widely accessible, especially in rural areas, and requires less imaging time and a lower cost (Abdul et al., 2021; Pereira et al., 2020). Given the absence of dependable automated toolkits, the demand for more diagnostic tools has increased. Accordingly, X-ray imaging scans are beneficial for a variety of diagnostics and procedures (Nayak et al., 2021). These scans aid in the noninvasive and painless diagnosis of a disease, therapy monitoring and surgical treatment planning (Kaplan, 2016). According to the Global Market Insights, it is anticipated that the worldwide medical X-ray market will attain a valuation of USD 16.5 billion by the year 2025 (Nkurunziza et al., 2022). The growing prevalence of cancer and other chronic diseases has stimulated the expansion of the medical X-ray market. Therefore, it is imperative to develop an automated and timely diagnostic system that can provide expeditious decisions and significantly mitigate diagnostic inaccuracies (Ravi et al., 2021).

Around the world, the COVID-19 epidemic has had catastrophic effects on people's health, economies, and societies. The COVID-19 virus continues to be a worry after vaccinations. As a result, various investigations on the early detection of COVID-19 have been carried out (Rahaman et al., 2020; Rubin et al., 2020). These investigations were conducted on X-ray images of COVID-19 patients in order to identify the critical elements for disease diagnosis. Additionally, radiologists noted that COVID-19-infected individuals exhibit chest radiograph abnormalities, suggesting that chest X-ray (CXR) images may disclose visual indicators associated with infection (Makris et al., 2020; Nour et al., 2020). The imaging tool is deemed to be a quick screening method for the instant identification of patients who may be epidemic suspects. Artificial intelligence (AI) methods, notably deep learning algorithms, have recently made the use of CXR images a practical screening tool (Bouchareb et al., 2021; Fusco et al., 2021; Nayak et al., 2021).

Furthermore, AI is one of the most effective computing environments that can assist the medical community in analysing abnormal symptoms and alerting patients and healthcare authorities (López-Cabrera et al., 2021). Applying relevant AI algorithms aid in the construction of a new system for the diagnosis and management of COVID-19 cases as well as quicker and more cost-effective decision-making (Ozturk et al., 2020). More so, by using medical imaging techniques including X-ray, CT, and magnetic resonance imaging scans of human body parts, AI can help in the diagnosis of infected patients (Caobelli, 2020). In light of this, a number of COVID-19 detections have been suggested in the academic literature and published based on different AI models (López-Cabrera et al., 2021; Nayak et al., 2021; Ozturk et al., 2020). Radiological imaging and cutting-edge AI algorithms may successfully identify this virus with high accuracy. The COVID-19 outbreak's quick spread made it necessary to have knowledge in this field, which raised interest in the creation of AI-based automated detection systems (Malik et al., 2021). Providing high-

quality medical care to all hospitals is difficult due to the existing shortage of radiologists. AI models that are simple, precise and quick can aid in immediately solving this challenge and assisting patients. Therefore, AI can also be beneficial in addressing the shortage of qualified healthcare providers in remote areas.

The utilisation of machine learning algorithms for automated medical diagnostics has become increasingly prevalent serving as a complementary tool for physicians and other healthcare practitioners (Shibly et al., 2020). A particular branch of AI technology, known as deep learning, has gained widespread recognition for its ability to produce comprehensive learning structures and algorithms, thereby yielding desired results from unprocessed data, without the need for manual feature extraction (Jain et al., 2021). Deep Transfer Learning (DTL), as defined by (Panigrahi et al., 2021), denotes the process of knowledge transfer from a well-trained source domain, possessing a substantial quantity of training samples, to a target domain with a relatively smaller set of training samples. Moreover, DTL involves the transfer of specific layers from a Convolutional Neural Network (CNN) model that has been pretrained on a vast number of images (Ahuja et al., 2021). The network architecture in this context generally excludes the fully connected classification output layer, retaining all the other layers for a new classification task.

Feature extraction in a CNN is responsible for gathering visual information from input images, which the neural network uses for subsequent classification. The final classification outcomes for the input image are contingent upon these extracted features (Liu, 2018). Numerous hybrid multi-deep transfer and machine learning (HMDTML) models, purposed for detecting and diagnosing COVID-19 cases from CXR images, have been presented in the literature, evaluated using various metrics (Alqudah et al., 2020; He et al., 2021; Minaee et al., 2020; Panwar et al., 2020; Pathan et al., 2021; Sethy et al., 2020; Toğaçar et al., 2020). Hence, further analysis is necessary to identify the existing HMDTML models and the evaluation metrics used to evaluate the proposed models in the literature. The study's originality can be further described as follows:

- The PSVNH-FWZIC method has been devised and established for the purpose of address the challenges posed by vagueness, uncertainty and hesitancy in ascertaining the evaluation criteria weights of the HMDTML models.
- A novel Multiple Criteria Decision Making (MCDM) ranking method named DLBD is proposed to manage dynamic changes with boundary value consequences and extended under a PSVNHF environment, to effectively address the challenges posed by vagueness, uncertainty and hesitancy in the benchmarking procedure.
- A new benchmarking process of HMDTML models is proposed by integrating PSVNH–FWZIC and PSVNH–DLBD with respect to the adopted decision matrix.

The structure of this manuscript is organized as follows. Section 2 encompasses a review of literature pertinent to this study. In Section 3, the research methodology employed to benchmark the HMDTML models for COVID-19 diagnosis, utilizing the PSVNH–FWZIC and PSVNH–DLBD methods is discussed. The outcomes and subsequent discussion pertaining to the benchmarking of HMDTML models are delineated in Section 4. Section 5 is dedicated to conducting sensitivity and comparative analyses, with the aim of assessing and validating the efficacy of the proposed method. The practical implication and limitations of the proposed method are discussed in Sections 6 and 7, respectively. Finally, Section 8 provides the conclusion of the manuscript.

2 Related Work

2.1 Research Gap, Challenges and Issues

As previously mentioned, the HMDTML model-based CXR images for detecting and diagnosing COVID-19 cases are available in the literature and evaluated on the basis of the different evaluation metrics (Alqudah et al., 2020; He et al., 2021; Minaee et al., 2020; Panwar et al., 2020; Pathan et al., 2021; Sethy et al., 2020; Toğaçar et al., 2020).

The utilization of CNN algorithms in the diagnosis of a range of diseases, as demonstrated by (Pathan et al., 2021), has inspired researchers to employ them in the context of COVID-19. (He et al., 2021) employed a COVID-19 CT dataset to evaluate and compare the efficacy of 3D and 2D CNN models. The authors of this study utilized four evaluation metrics, namely accuracy, precision, sensitivity, and F1-score, to rank the models. (Alqudah et al., 2020) introduced a hybrid framework that incorporated CNN and machine learning techniques, specifically support vector machine (SVM) and k-nearest neighbours (kNN), for the purpose of detecting COVID-19 disease through the analysis of CXR images. This study employed six evaluation metrics, namely accuracy, sensitivity, specificity, precision, F1-score, and Matthews correlation coefficient. In their study, (Sethy et al., 2020) utilized a combination of 13 CNN models and SVM to extract intricate features from X-ray images and subsequently detect cases of COVID-19. The evaluation of the integrated models was conducted based on four metrics, specifically accuracy, sensitivity, false-positive rate, and F1-score. (Toğaçar et al., 2020) used a transfer learning approch to train deep learning models, namely MobileNetV2 and SqueezeNet, in conjunction with SVM on the dataset of CXR images for COVID-19 detection. The authors employed sensitivity, F1-score, specificity, precision, and accuracy metrics for the assessment of their proposed models. In their study, (Minaee et al., 2020) used DTL to train several models including DenseNet-121, SqueezeNet, ResNet18, and ResNet50 on the CXR COVID-19 dataset. The authors then assessed the performance of these models using sensitivity and specificity metrics. In their study, (Panwar et al., 2020) trained a deep learning neural network named

nCOVnet on the CXR COVID-19 dataset. The performance of the model was assessed using metrics such as accuracy and the area under the curve (AUC).

Another group of researchers utilized optimization algorithms to effectively identify the optimal combination of hyperparameters for deep CNNs. In (Cai et al., 2023; Hu et al., 2021), the Chimp optimisation algorithm was employed to train the fully connected layers of deep CNNs for Covid19 diagnosis using CXR images from the COVID-Xray-5k and COVIDetectionNet datasets. The study authors employed seven evaluation metrics, including standard deviation (STD), sensitivity, specificity, accuracy, AUC, precision-recall, and processing time, to validate their model. In the studies proposed by (Saffari et al., 2022; Wang et al., 2022) whale optimisation algorithm was used, in conjunction with a fuzzy system, which utilised to train the fully connected layers of deep CNNs for the purpose of diagnosing Covid-19 from CXR images sourced from the COVID-Xray-5k dataset. The study authors employed seven evaluation metrics, including accuracy, processing time, STD, ROC, precision-recall, and F1-Score, to validate their model. The study of (Khishe et al., 2021) utilized a heuristic optimization technique to efficiently determine the optimal combination of hyperparameters for CNNs. This framework began by optimizing a foundational CNN, serving as the initial stage of the evolutionary process. Afterwards, a maximum of two extra convolutional layers are incrementally incorporated into the existing convolutional structure during an additional optimization phase. The model was trained using COVID-19 CXR images obtained from the COVID X-ray-5k dataset. The researchers used six evaluation metrics, namely accuracy, processing time, STD, ROC, and precision-recall, to validate their model. Despite the remarkable achievements of these works, there has been a lack of proposing dynamic decision matrix that allows for benchmarking the models against others by considering all matrices simultaneously. Furthermore, the importance of these matrices was not taken into consideration. According to the preceding discussion, different deep learning and machine learning models were presented and evaluated using multiple criteria as individual processes, which is considered a research gap.

The integration of MCDM methods was employed in a study conducted by (M. Mohammed et al., 2020) to rank the HMDTML using (Nanni et al., 2017) datasets. The authors employed TOPSIS and entropy to compare the performance of machine learning models trained on the COVID-19 CXR dataset. The assessment metrics utilized for these models encompassed AUC, precision, F1-score, recall, accuracy, true-positive (TP), true-negative (TN), false-positive (FP), false-negative (FN), and time. This study was found to have a limitation in its use of recall and TP as evaluation metrics, as they share similarities in their approach and mathematical equations. Furthermore, the TOPSIS technique is confronted with three fundamental obstacles, namely the utilization of Euclidean distance, normalization techniques, and the absence of a mechanism to determine the relative importance of criteria. (Alamleh et al., 2022; Ramleh,

Yatim, et al., 2022). Accordingly, the entropy method employed in the study conducted by (M. Mohammed et al., 2020) for resolving the weighting problem pertaining to criteria importance is an objective method that relies on data distribution while disregarding the experts' opinion of the criteria. This approach renders the weighting process unreasonable. Thus, there is a need to employ human approaches to address this issue.

In 2022, a novel method known as fuzzy-weighted zero-inconsistency (FWZIC) (R. T. Mohammed et al., 2022) was introduced and is regarded as the most effective subjective approach for estimating the weights of evaluation criteria. The initial iteration of FWZIC, as presented by R. T. Mohammed et al. (2022), employed triangular fuzzy numbers (TFNs). However, this approach was deemed impractical for two reasons. Firstly, the definition of membership values in TFNs was found to be difficult (Liao, 2015). Secondly, TFNs were unable to effectively account for and address the impact of data uncertainty (Mathew et al., 2020). The vagueness, uncertainty and hesitancy concerns have a great influence on the ultimate evaluation of the alternatives within the framework of MCDM (Dalic et al., 2020; Deveci, Pamucar, & Delen, 2023; Zaidan et al., 2023). These concerns commonly arise from the viewpoints of individuals in positions of authority who make decisions. Therefore, reasonably assessing and benchmarking HMDTML models trained on COVID-19 CXR images remains an open challenge from the theoretical perspective.

2.2 Theoretical Justifications

Salih et al., (2020) introduced a novel MCDM approach, referred to as the fuzzy decision by opinion score method (FDOSM), which utilizes the closeness of the ideal solution for each criterion to the other values within that criterion as a means of addressing the challenge of measuring distance. This method effectively resolves discrepancies that arise from the use of various normalising techniques. Thus, this particular method effectively addresses the primary limitations associated with mathematical methodologies employed in the TOPSIS method. The utilisation of expert guidance in conjunction with an opinion matrix facilitates the streamlining of the decision-making process by determining the ideal value for each criterion. Moreover, this method compares the ideal value of a certain criterion to the other values of each alternative under that criterion. The method proposed by Salih et al., (2020) exhibits superiority due to several reasons: (i) minimising the processing time required for comparisons; (ii) avoiding comparing unrelated criteria (unnatural comparisons); (iii) tackling the problem of inconsistency, which is the primary issue with subjective methods; (iv) solving the missing values problem; (v) implicitly estimating the criteria weights; and (vi) solving the problem of immeasurable values. In addition, FDOSM addressed the vagueness, uncertainty and hesitancy of information by employing fuzzy sets (FSs) including but not limited to q-rung orthopair fuzzy rough sets (Deveci et al., 2023), Pythagorean FSs (Rozi et al., 2023), and fermatean probabilistic hesitant FS (Zaidan et al., 2023). Notwithstanding its benefits, FDOSM is impeded

by two inherent deficiencies.

First, FDOSM lacks the ability to assign precise weights for assessing criteria in terms of their significance and necessitates the use of an external approach to fulfil this requirement (Rozi et al., 2022; Baqer et al., 2023). Accordingly, the amalgamation of FDOSM and FWZIC has been employed by researchers to tackle diverse intricate MCDM problems in the literature (Rozi et al., 2022; Z. K. Mohammed et al., 2023). As previously mentioned, FWZIC exhibits an inability to effectively process information that is characterised by vagueness, uncertainty, and hesitancy. Therefore, FWZIC has been extended under to various classical (e.g. TFN (R. T. Mohammed et al., 2022) and trapezoidal fuzzy numbers) and traditional (e.g. q-rung orthopair fuzzy rough sets, neutrosophic cubic FSs, Pythagorean FSs, interval-valued spherical fuzzy rough sets, and fermatean probabilistic hesitant FS) environments (Qumar et al., 2022). In addition to addressing the vagueness, ambiguity and hesitancy of information, all of these traditional fuzzy environments are used with FWZIC to solve issues experienced with earlier versions of FWZIC that employed classical fuzzy environments. Despite the considerable efforts made, the unresolved issues of informational vagueness, ambiguity, and hesitancy persist.

Second, the original version of FDOSM has been used on subjective and objective evaluation matrices and tested on communication and sport case studies (Dalic et al., 2020). However, subjectively evaluating objective values is unreasonable. Consequently, FDOSM is incapable of handling a dynamic change with the consequences of boundary values. As previously mentioned, FDOSM's methodology relies on the selection of the ideal solution for each criterion and the comparison of that ideal to other values under that criterion, with the assistance of experts. Accordingly, subjectively comparing and evaluating dynamic measured values are impractical. For instance, if C1 is a benefit criterion with boundary values between 0 and 100, then selecting the optimal solution, which is 100, does not require the opinion of experts. Thus, a new MCDM method capable of handling dynamic change with boundary value consequences must be developed.

In this work, a new method named dynamic localisation-based decision (DLBD) which consists of three main steps has been developed. This method generates a dynamic localisation decision matrix based on the upper and lower boundaries of the scale and the length of the scale. This method can handle a dynamic change with the consequences of boundary values. Considering the advantages of FDOSM, the DLBD method can reduce the variance of outcomes across normalisation methods. Additionally, this method decreases the processing time required to perform localisation. Furthermore, this method uses upper and lower boundaries to overcome the distance measurement problem. The DLBD method can also automatically estimate the implicit weights for the set of values in each criterion by using scale values. However, the DLBD method still requires an external method to estimate the explicit weight of the criteria.

Accordingly, this mechanism is integrated with FWZIC in this study to evaluate and benchmark HMDTML models trained on COVID-19 CXR images based on seven evaluation criteria. The most important contribution of this work is the enrichment of the entire field of decision-making through the development of a novel method that will aid experts in addressing complex MCDM problems.

FSs including FDOSM (Rozi et al., 2022; Salih et al., 2020) are necessary for DLBD to handle the ambiguity, uncertainty and hesitancy of information when the dynamic localisation process has been performed within the lower and upper boundaries for each scale. Several novel forms of FS theory have been formulated, including but not limited intuitionistic FSs, hesitant FSs, dual hesitant FSs, interval-valued intuitionistic FSs, and dual hesitant fuzzy probability. Incorporating statistical uncertainty into actual productions is imperative. The efficacy of the probabilistic method may be limited in addressing epistemic uncertain problems. Accordingly, these issues motivate scholars to integrate FS and probabilistic theories to create a new fuzzy concept. Shao et al. (2018) proposed the concept of probabilistic single-valued neutrosophic hesitant FS (PSVNHFS) based on hesitant FS, probabilistic dual hesitant FS, neutrosophic FS and interval neutrosophic hesitant FS to solve MADM problems. The authors integrated the SVNHFS and probability information by concurrently expressing the membership degree values of truth, indeterminacy, and falsity alongside their corresponding probability values. The PSVNHFS provides additional information to aid in the decision-making process (Şahin & Altun, 2020).

To the authors' best knowledge, the existing MCDM ranking methods are unable to handle the dynamic change with consequences of boundary values of the HMDTML models. In addition, informational vagueness, uncertainty and hesitancy issues remain unsolved in the MCDM ranking and weighting methods. Accordingly, this study proposed and developed a novel combined DLBD with FWZIC under a PSVNHFS environment to benchmark the HMDTML models. The novel DLBD method is proposed to generate a dynamic localisation decision matrix based on the upper and lower boundaries and the length of the scale. The superiority of DLBD derives from its ability to manage dynamic changes with boundary value consequences. In addition, the employment of PSVNHFS with DLBD and FWZIC could successfully overcome the vagueness, uncertainty and hesitancy of information issues in the benchmarking process. Given the aforementioned rationales and benefits associated with utilising PSVNH-FWZIC and PSVNH-DLBD as means of ascertaining criteria weight and establishing benchmarks for alternatives, we are motivated to suggest a novel fuzzy MCDM approach for resolving the issue pertaining to the selection of HMDTML models.

3 Developing a Deep Transfer and Machine Learning Models

The developing a deep transfer and machine learning models applied in this study is illustrated in Fig. 1. Section 3.1 outlines the construction of the decision matrix for the HMDTML model. The application of

the COVID-19 case study is presented in Section 3.2. This section is divided into two subsections. Section 3.2.1 presents the PSVNH-FWZIC method's developmental process, which is utilized for the purpose of assigning weights to the performance evaluation criteria of the HMDTML models. Section 3.2.2 presents development of the PSVNH–DLBD ranking method.

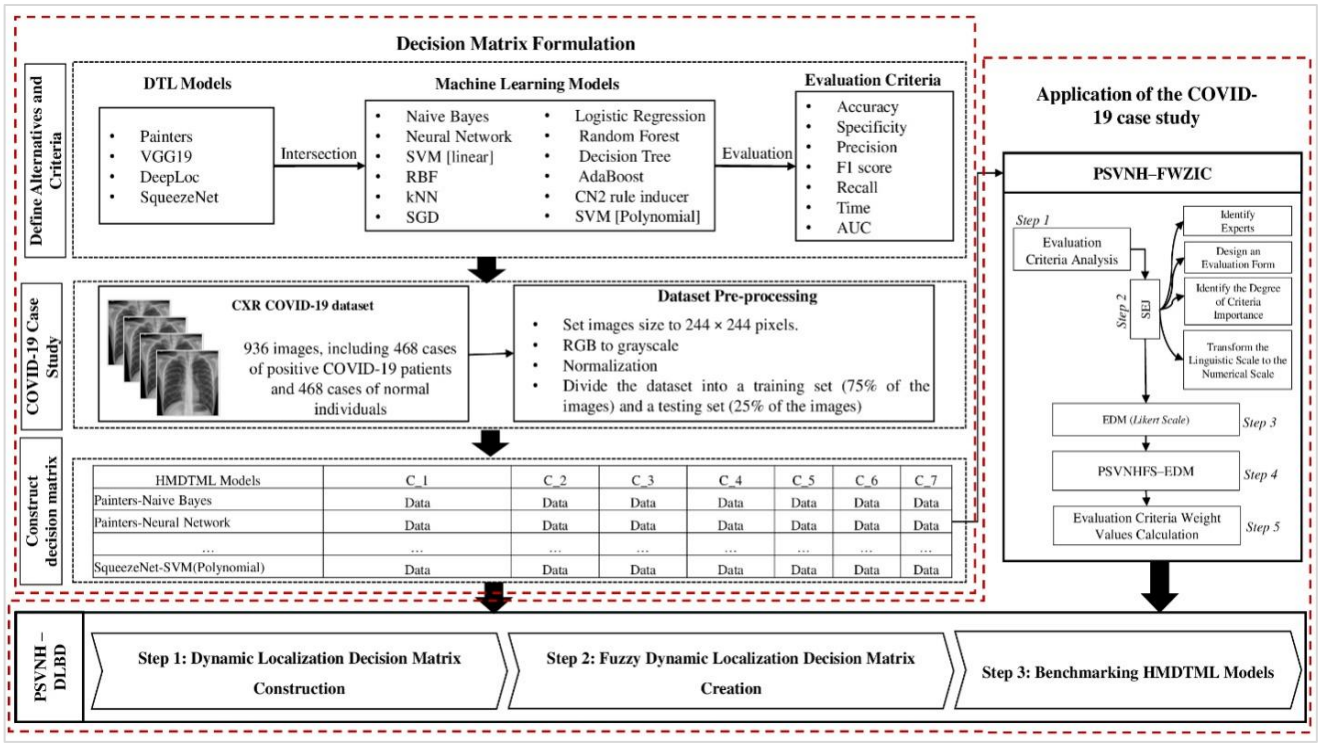


Fig. 1. Research methodology.

3.1 Decision Matrix Formulation

The formulation of the decision matrix is summarised in the following four steps:

Step 1: Create and define alternatives

(M. Mohammed et al., 2020) integrated a single DTL model (e.g. Inception v3) with twelve supervised machine learning algorithms (e.g. naive Bayes, neural network, SVM [linear], radial basis function [RBF], kNN, stochastic gradient descent [SGD], logistic regression, random forest, decision tree, AdaBoost, CN2 rule inducer and SVM [Polynomial]). In this study, the alternatives are created by integrating four DTL models, Painters, VGG19, DeepLoc and SqueezeNet, which are well-known and widely used in different studies (Yadav et al., 2022) with the same set of supervised machine-learning algorithms. The process of extracting valuable features from unprocessed input data is an essential step in acquiring precise and meaningful representations. In addition, DTL is a technique that enables the training of data at a lower calculation cost and with fewer datasets. The transfer learning technique is used to transfer the information

obtained by a pretrained model on a large dataset to the model to be trained on a new dataset (Khan & Aslam, 2020). Deep learning models are used to automatically extract features to tackle a certain problem. The features extracted by deep learning, known as learned features, are incredibly effective. Models that extract and classify features using deep learning outperform models that classify manually derived features by a large margin (Nanni et al., 2017). VGG19 and SqueezeNet are pretrained by an ImageNet dataset. However, DeepLoc and Painters are pretrained using images of yeast cells and artwork, respectively. These models are most frequently used in numerous deep-learning image classification problems. Moreover, the selected models are preferable because they are simple to implement and computationally less expensive, making them excellent building blocks for learning purposes. In addition, the twelve supervised machine learning algorithms are applied to improve feature detection and to match the new HMDTML model with the most exhaustive and thoroughly researched algorithms. The selected algorithms are feasible due to their high accuracy and adaptability for achieving superior outcomes. In (M. Mohammed et al., 2020) is provided further details about each of the employed supervised machine learning algorithms. Forty-eight HMDTML models are obtained, as shown in Fig. 2.

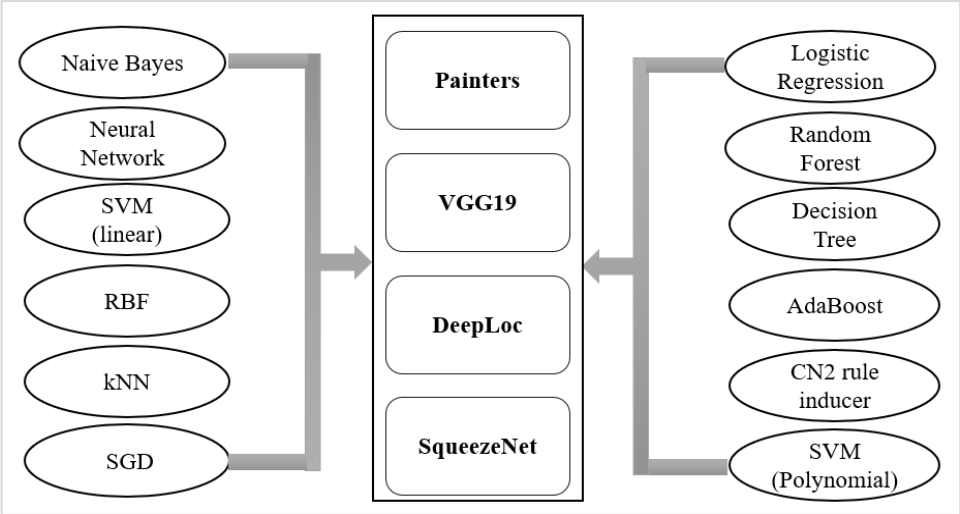


Fig. 2. Intersection process between four DTL models with twelve supervised machine learning algorithms.

Step 2: Define and identify performance evaluation criteria

The performance measurements are essential in determining how the proposed HMDTML models effectively fulfil the objective. (M. Mohammed et al., 2020) evaluated their HMDTML models using ten metrics as mentioned in Section 2.1. In this study, seven metrics are used to evaluate the developed HMDTML models: AUC (C_1), accuracy (C_2), F1-score (C_3), precision (C_4), recall (C_5), time (C_6) and specificity (C_7). Typically, due to, TP, TN, FP and FN are used to formulate these metrics. Accordingly,

the comparison between the different models on those four typical options is considered implicitly, and there is no need to consider them in the decision matrix. In addition, specificity is used in this study, as it can be calculated from TP and TN to test the probability of a negative test, conditioned on truly being negative. The term "TP" denotes the count of accurately labelled positive samples, while "TN" signifies the count of accurately identified negative samples. "FP" refers to the negative samples that were inaccurately identified as positive, and "FN" represents the positive samples that were predicted as negative. The equations for deriving the values of these metrics are shown in (1), (2), (3), (4), (5) and (6) (Goel et al., 2021; M. Mohammed et al., 2020).

$$C_2 = \frac{TP+TN}{TP+FP+TN+FN} \quad (1)$$

$$C_3 = 2 \times \frac{C_4 \times C_2}{C_4 + C_2} \quad (2)$$

$$C_4 = \frac{TP}{TP+FP} \quad (3)$$

$$C_5 = \frac{TP}{TP+FN} \quad (4)$$

$$C_6 = To - Ti, \quad (5)$$

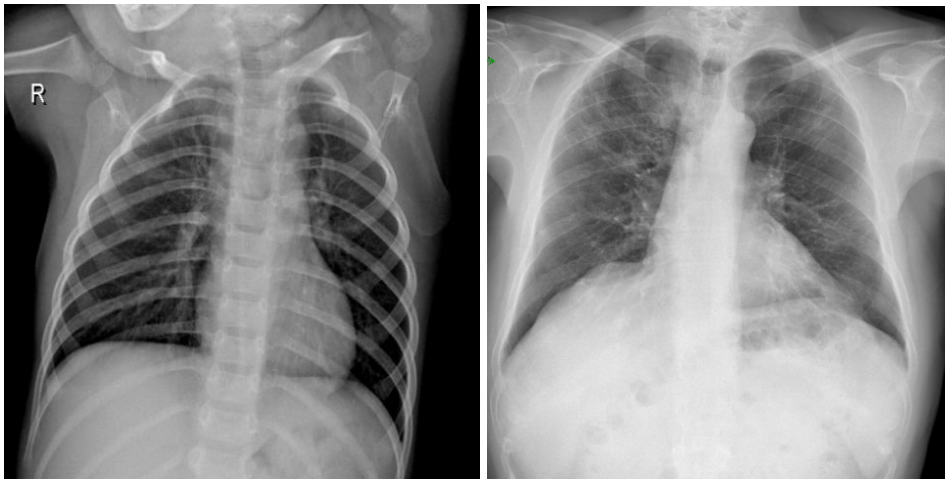
where To represents the time of processing to obtain outputs, and Ti represents the sample input.

$$C_7 = \frac{TN}{(TN+TP)} \quad (6)$$

Step 3: Case study

Datasets selection. The present study employs the proposed methods to a case study on COVID-19 as a means of demonstrating their feasibility. As mentioned in the Introduction section, there is only one study (M. Mohammed et al., 2020) used MCDM methods as an integrated process to rank the HMDTML based on (Goel et al., 2021) dataset. However, the dataset exclusively comprises individuals who have tested positive for COVID-19, as well as those who have tested negative for COVID-19, including cases of MERS, SARS, and ARDS. In the present study, the training of HMDTML is based on the individuals who have tested positive for COVID-19 and non-infected individuals. Consequently, in accordance with the studies conducted by (Loey et al., 2020; M. Mohammed et al., 2020), two publicly available datasets were selected as the main sources of CXR images. The 1st dataset (Cohen et al., 2020) contains 660 CXR images, 468 of which depict COVID-19-positive patients and the remaining depict COVID-19-negative patients (including MERS, SARS and ARDS). It should be noted that this study exclusively enrolled individuals who tested positive for COVID-19 and those who were deemed healthy. Additional instances of pneumonia caused by viruses and bacteria were eliminated from consideration. Thus, the CXRs of the healthy individuals were

not used from this dataset. The 2nd dataset (Kermany et al., 2018) consists of 5679 CXR images divided into two categories: normal images with 1406 samples and pneumonia images with 4273 samples. A total of 468 out of 1406 images were selected from the normal category. To address the data imbalance, the images are evenly distributed between the two categories, namely positive and non-infected COVID-19 individuals. The total collection had 936 images, including 468 cases of positive COVID-19 patients and 468 cases of normal individuals. Fig. 3 shows the CXR image samples of a patient infected with COVID-19 and normal cases extracted from the selected datasets.



(a) (b)

Fig. 3. CXR images (a) Normal and (b) COVID-19.

Datasets Preprocessing. The following is a description of the preprocessing steps employed in this study. First, the CXR images had different sizes. Therefore, these images were resized to 224 × 224 pixels. Second, all images are converted from RGB to greyscale. Third, data normalisation was conducted for improved model learning. Fourth, binary classification is employed because the dataset is divided into two categories: normal and COVID-19. Finally, the dataset is divided into a training set (75% of the images) and a testing set (25% of the images). The resulting dataset is fed into the HMDTML models for training.

Step 4: Construct a uniform dynamic decision matrix

The decision matrix encompasses an important part of HMDTML model benchmarking. The 48 HMDTML models (M1, M2, …, M48) represent the alternatives in the decision matrix, whereas the seven predefined metrics represent the evaluation criteria. The formulation of the uniform dynamic decision matrix is based on the crossover between the 48 HMDTML models and the seven evaluation criteria. Table 1 presents the formulation of the uniform dynamic decision matrix.

Table 1
Uniform dynamic decision matrix of HMDTML model formulation.

	HMDTML Models/Criteria	C_1	C_2	C_3	C_4	C_5	C_6	C_7
M1	Painters-Naive Bayes	…	…	…	…	…	…	…
M2	Painters-Neural Network	…	…	…	…	…	…	…
M3	Painters-SVM (linear)	…	…	…	…	…	…	…
M4	Painters-RBF	…	…	…	…	…	…	…
M5	Painters-kNN	…	…	…	…	…	…	…
M6	Painters-SGD	…	…	…	…	…	…	…
M7	Painters-Logistic Regression	…	…	…	…	…	…	…
M8	Painters-Random Forest	…	…	…	…	…	…	…
M9	Painters-Decision Tree	…	…	…	…	…	…	…
M10	Painters-AdaBoost	…	…	…	…	…	…	…
M11	Painters-CN2 rule inducer	…	…	…	…	…	…	…
M12	Painters-SVM (Polynomial)	…	…	…	…	…	…	…
M13	VGG19-Naive Bayes	…	…	…	…	…	…	…
M14	VGG19-Neural Network	…	…	…	…	…	…	…
M15	VGG19-SVM (linear)	…	…	…	…	…	…	…
M16	VGG19-RBF	…	…	…	…	…	…	…
M17	VGG19-kNN	…	…	…	…	…	…	…
M18	VGG19-SGD	…	…	…	…	…	…	…
M19	VGG19-Logistic Regression	…	…	…	…	…	…	…
M20	VGG19-Random Forest	…	…	…	…	…	…	…
M21	VGG19-Decision Tree	…	…	…	…	…	…	…
M22	VGG19-AdaBoost	…	…	…	…	…	…	…
M23	VGG19-CN2 rule inducer	…	…	…	…	…	…	…
M24	VGG19-SVM (Polynomial)	…	…	…	…	…	…	…
M25	DeepLoc-Naive Bayes	…	…	…	…	…	…	…
M26	DeepLoc-Neural Network	…	…	…	…	…	…	…
M27	DeepLoc-SVM (linear)	…	…	…	…	…	…	…
M28	DeepLoc-RBF	…	…	…	…	…	…	…
M29	DeepLoc-kNN	…	…	…	…	…	…	…
M30	DeepLoc-SGD	…	…	…	…	…	…	…
M31	DeepLoc-Logistic Regression	…	…	…	…	…	…	…
M32	DeepLoc-Random Forest	…	…	…	…	…	…	…
M33	DeepLoc-Decision Tree	…	…	…	…	…	…	…
M34	DeepLoc-AdaBoost	…	…	…	…	…	…	…
M35	DeepLoc-CN2 rule inducer	…	…	…	…	…	…	…
M36	DeepLoc-SVM (Polynomial)	…	…	…	…	…	…	…
M37	SqueezeNet-Naive Bayes	…	…	…	…	…	…	…
M38	SqueezeNet-Neural Network	…	…	…	…	…	…	…
M39	SqueezeNet-SVM (linear)	…	…	…	…	…	…	…
M40	SqueezeNet-RBF	…	…	…	…	…	…	…
M41	SqueezeNet-kNN	…	…	…	…	…	…	…
M42	SqueezeNet-SGD	…	…	…	…	…	…	…

M43	SqueezeNet-Logistic Regression	…	…	…	…	…	…	…
M44	SqueezeNet-Random Forest	…	…	…	…	…	…	…
M45	SqueezeNet-Decision Tree	…	…	…	…	…	…	…
M46	SqueezeNet-AdaBoost	…	…	…	…	…	…	…
M47	SqueezeNet-CN2 rule inducer	…	…	…	…	…	…	…
M48	SqueezeNet-SVM (Polynomial)	…	…	…	…	…	…	…

3.2 *Application of the COVID-19 case study*

The data of the uniform dynamic decision matrix were numerically obtained from the evaluation criteria representing an objective process. All evaluation criteria were beneficial (the height value was the best) such as AUC, accuracy, F1-score, precision, recall and specificity, except that the time was costly (the lower value was the best). Evaluating and benchmarking HMDTML models are challenging if not impossible from a human perspective. This process is also difficult due to three primary issues (M. Mohammed et al., 2020): multiple evaluation criteria, data variance and importance of criteria and trade-offs between various criteria. However, the existing MCDM ranking methods are unable to handle the dynamic change with the consequences of boundary values of the HMDTML models. In addition, informational vagueness, uncertainty and hesitancy issues remained unsolved in the MCDM ranking and weighting methods. Accordingly, the proposed study comes out with the solution. The application of the proposed method to the COVID-19 case study is performed as follows:

i. The preprocessed dataset is fed to 48 HMDTML models for training them on COVID-19-positive or COVID-19-negative (normal) cases.
ii. The performance of the 48 HMDTML models is measured using the seven preselected metrics (i.e. AUC, accuracy, F1-score, precision, recall, time and specificity). The collected data represent the values of the uniform dynamic decision matrix.
iii. The developed PSVNH–FWZIC method is applied to the performance evaluation criteria to determine their importance level (Section 3.2.1)
iv. The proposed PSVNH–DLBD method is applied to the uniform dynamic decision matrix to benchmark HMDTML models (Section 3.2.2).

3.2.1 PSVNH–FWZIC Weighting Method Formulation

In the context of MCDM, FWZIC (R. T. Mohammed et al., 2022) is a new subjective method for estimating the relative importance of criteria with zero inconsistencies. In many real-world applications, input data cannot be precisely measured. Factors such as indirect measurements, model estimation, subjective interpretation, decision-makers' (experts') judgement and the availability of information from multiple sources can contribute to imprecision and ambiguity.The precise weight that should be associated

with a certain criterion is difficult to identify in the presence of uncertainty. The FWZIC supports decision-makers in overcoming their own subjectivity when estimating the importance level of the criteria. Therefore, FWZIC is developed to solve these issues and estimate the criteria weights of the HMDTML models under the PSVNHFS environment. The steps of PSVNH–FWZIC formulation and development are as follows:

Step 1: Evaluation Criteria Analysis

The set of evaluation criteria used for evaluating and benchmarking the HMDTML models is analysed and examined.

Step 2: Structured expert judgement (SEJ)

In this step, four sequential substeps are needed: (i) experts with prior knowledge of the study's subject are identified and recruited. Three experts were recruited to evaluate the evaluation criteria. (ii) A form of evaluation is created and subsequently approved by the enlisted experts. This form has undergone testing and received endorsement from the three experts who were enlisted in the preceding step. (iii) The experts used five linguistic expressions (e.g. very important [VI], important [I], average [Av], low important [LI] and very low important [VLI]) to assign the degrees of importance to all criteria (see Table 2). These terms denote the opinions of experts, which are gathered through the evaluation form. (iv) The collected data (linguistic expressions) are replaced with their equivalent numerical scale (see Table 2).

Table 2
Linguistic expressions, numerical scale and PSVNHFNs.

Linguistic expressions	Numerical scale/scale values	PSVNHFNs $T(x)\|P^T(x)$				$I(x)\|P^I(x)$				$F(x)\|P^F(x)$			
		α_1	$P^T_{\alpha_1}$	α_2	$P^T_{\alpha_2}$	β_1	$P^I_{\beta_1}$	β_2	$P^I_{\beta_2}$	γ_1	$P^I_{\gamma_1}$	γ_2	$P^I_{\gamma_2}$
VI	1	0.95	0.8	0.9	0.2	0.05	0.8	0.1	0.2	0.05	0.8	0.1	0.2
I	2	0.75	0.7	0.7	0.3	0.25	0.7	0.3	0.3	0.25	0.7	0.3	0.3
AV	3	0.55	0.5	0.5	0.5	0.45	0.5	0.5	0.5	0.45	0.5	0.5	0.5
LI	4	0.35	0.7	0.25	0.3	0.65	0.7	0.75	0.3	0.65	0.7	0.75	0.3
VLI	5	0.15	0.8	0.1	0.2	0.85	0.8	0.9	0.2	0.85	0.8	0.9	0.2

Step 3: Establishment of Expert Decision Matrix (EDM)

In this step, the EDM is constructed. The primary constituents of the EDM consist of the evaluation criteria for the HMDTML models and the experts involved in the process. The experts' evaluations of the m-criteria $\{C_i, i = 1, \dots, m\}$ form the EDM matrix, as shown in Eq. (7). In the proposed method, the EDM serves as the basis for further analysis.

$$\text{EDM} = \begin{bmatrix} E_{11} & \cdots & E_{1m} \\ \vdots & \ddots & \vdots \\ E_{l1} & \cdots & E_{km} \end{bmatrix}, \tag{7}$$

where E_{ij} is the numerical scale corresponding to the linguistic expression (VI, I, AI, LI, VLI) given by the j-th expert to the i-th criterion.

Step 4: Application of PSVNHFS

PSVNHFS theory is applied on the produced EDM to create PSVNHFS–EDM as follows:

$$\widetilde{\text{EDM}} = \begin{bmatrix} \tilde{E}_{11} & \cdots & \tilde{E}_{1m} \\ \vdots & \ddots & \vdots \\ \tilde{E}_{l1} & \cdots & \tilde{E}_{km} \end{bmatrix} \tag{8}$$

In this context, all numerical values within the EDM are replaced with their associated probabilistic single-valued neutrosophic hesitant fuzzy numbers (PSVNHFNs) given in Table 2. The robustness of PSVNHFS can be traced back to its ability to cope with complex and ambiguous data. The description of PSVNHFS and PSVNHFN are provided in Definition 1.

Definition 1. (Shao et al., 2018) *Let X be a fixed set. A PSVNHFS on X is defined as follows:*

$NP = \{\langle x, T(x)|P^T(x), I(x)|P^I(x), F(x) \mid P^F(x)\rangle \mid x \in X\}$.

The potential elements are represented in three distinct components, denoted as $T(x)|P^T(x), I(x)|P^I(x), F(x) \mid P^F(x)$. $T(x)$, $I(x)$ and $F(x)$ are finite subsets of $[0,1]$ representing the hesitant degrees of possible truth, indeterminacy, and falsity of x with respect to the set X. The corresponding probabilistic information for the aforementioned three categories of degrees are represented by $P^T(x)$, $P^I(x)$ and $P^F(x)$, which are also subsets of $[0,1]$ and having the same cardinality as their associated degrees sets. For $\alpha_a \in T(x), \beta_b \in I(x)$, and $\gamma_c \in F(x)$, the following conditions are satisfied:

$$P^T_{\alpha_a} \in P^T(x), P^I_{\beta_b} \in P^I(x), P^F_{\gamma_c} \in P^F(x); \sum_{i=1}^{|T(x)|} P^T_{\alpha_i} \le 1, \sum_{i=1}^{|I(x)|} P^I_{\beta_i} \le 1, \sum_{i=1}^{|F(x)|} P^F_{\gamma_i} \le 1,$$

and

$$0 \le \alpha^+ + \beta^+ + \gamma^+ \le 3,$$

where $\alpha^+ = max\{T(x)\}$, $\beta^+ = max\{I(x)\}$ and $\gamma^+ = max\{F(x)\}$; and $|.|$ denotes the cardinality of a set.

An element in NP is called a PSVNHFN and is represented as:

$$\langle \left(\alpha_1 \middle| P^T_{\alpha_1}, \ldots \alpha_{|T(x)|} \middle| P^T_{\alpha_{|T(x)|}}\right), \left(\beta_1 \middle| P^I_{\beta_1}, \ldots \beta_{|I(x)|} \middle| P^I_{\beta_{|I(x)|}}\right), \left(\gamma_1 \middle| P^F_{\gamma_1}, \ldots \gamma_{|F(x)|} \middle| P^F_{\gamma_{|F(x)|}}\right) \rangle,$$

for $x \in X$. For convenience, hereafter a PSVNHFN is denoted by $N = \langle \mathrm{T} \middle| P^{\mathrm{T}}, I \middle| P^{\mathrm{I}}, \mathrm{F} \mid P^{\mathrm{F}} \rangle$.

Step 5: Calculation of Evaluation Criteria Weight

This step calculates the final weights of the evaluation criteria $(w_1, w_2, \ldots, w_m)^T$ as follows:

(i) The fuzzy weights of the evaluation criteria are determined by aggregating the PSVNHFNs present in the PSVNHFS-EDMs of the l experts who were enlisted for the task. The probabilistic single-valued neutrosophic hesitant fuzzy weighted averaging (PSVNHFWA) operator (Shao et al., 2018) shown in Eq. (9) is modified, and a probabilistic single-valued neutrosophic hesitant fuzzy averaging (PSVNHFA) operator shown in Eq. (10) is used for the aggregation.

Let $N_j = \langle T_j \left| P^{T_j}, I_j \right| P^{I_j}, F_j \mid P^{F_j} \rangle$, for $j = 1, \ldots, r$, be PSVNHFNs. Then, the PSVNHFWA operator is defined as follows:

$$\text{PSVNHFWA}\ (N_1, N_2, \ldots, N_r) = w_1 N_1 \oplus w_2 N_2 \oplus \ldots . \oplus w_r N_r,$$

where $w = (w_1, w_2, \ldots . , w_r)^T$ represents the weights vector with $\sum_{j=1}^{r} w_j = 1$. Subsequently, the outcome of aggregation utilising PSVNHFWA can be achieved as follows:

$$\text{PSVNHFWA}\ (N_1, N_2, \ldots, N_r) =$$

$$\left(\begin{array}{c} \bigcup_{(\alpha_j)_{j=1,\ldots,r} \in T_1 \times T_2 \times \ldots \times T_r} \left(1 - \prod_{j=1}^{r} \left(1 - \alpha_j\right)^{w_j} \mid \Pi_{j=1}^{r} P_{\alpha_j}^{T_j} \right), \\ \bigcup_{(\beta_j)_{j=1,\ldots,r} \in I_1 \times I_2 \times \ldots \times I_r} \left(\prod_{j=1}^{r} \left(\beta_j\right)^{w_j} \mid \Pi_{j=1}^{r} P_{\beta_j}^{I_j} \right) , \bigcup_{(\gamma_j)_{j=1,\ldots,r} \in F_1 \times F_2 \times \ldots \times F_r} \left(\prod_{j=1}^{r} \left(\gamma_j\right)^{w_j} \mid \Pi_{j=1}^{r} P_{\gamma_j}^{F_j} \right) \end{array} \right). \tag{9}$$

In particular, if $w = \left(\frac{1}{r}, \frac{1}{r}, \cdots, \frac{1}{r}\right)^T$, then the PSVNHFWA operator is reduced to the PSVNHFA operator as follows:

$$\text{PSVNHFA}\ (N_1, N_2, \ldots, N_r) =$$

$$\left(\begin{array}{c} \bigcup_{(\alpha_j)_{j=1,\ldots,r} \in T_1 \times T_2 \times \ldots \times T_r} \left(1 - \prod_{j=1}^{r} \left(1 - \alpha_j\right)^{1/r} \mid \Pi_{j=1}^{r} P_{\alpha_j}^{T_j} \right), \\ \bigcup_{(\beta_j)_{j=1,\ldots,r} \in I_1 \times I_2 \times \ldots \times I_r} \left(\prod_{j=1}^{r} \left(\beta_j\right)^{\frac{1}{r}} \mid \Pi_{j=1}^{r} P_{\beta_j}^{I_j} \right) , \bigcup_{(\gamma_j)_{j=1,\ldots,r} \in F_1 \times F_2 \times \ldots \times F_r} \left(\prod_{j=1}^{r} \left(\gamma_j\right)^{1/r} \mid \Pi_{j=1}^{r} P_{\gamma_j}^{F_j} \right) \end{array} \right). \tag{10}$$

The aggregation of the PSVNHFNs present in the PSVNHFS-EDMs in (8) is then $\bar{E}_j = PSVNHFA\left(\tilde{E}_{1j}, \tilde{E}_{2j}, \ldots, \tilde{E}_{lj}\right)$, for $j = 1, \ldots, m$.

(ii) The evaluation criteria's fuzzy weights are transformed into crisp weights through the utilisation of the PSVNHFS score function (Shao et al., 2018), as shown in Eq. (11). For any PSVNHFN $N = \langle T | P^T, I | P^I, F | P^F \rangle$, a score function is defined as

$$s(N) = \frac{\left(\frac{1}{|T|}\sum_{\alpha \in T} \left(\alpha \cdot P_{\alpha}^{T}\right)\right) + \left(\frac{1}{|I|}\sum_{\beta \in I} \left(1-\beta\right) \cdot P_{\beta}^{I}\right) + \left(\frac{1}{|F|}\sum_{\gamma \in F} \left(1-\gamma\right) \cdot P_{\gamma}^{F}\right)}{3}. \tag{11}$$

(iii) The criteria's cumulative weights are set to one. If this condition is not satisfied, then the weights are rescaled using Eq. (12).

$$w_j = s\left(\bar{E}_j\right) / \sum_{j=1}^{m} s\left(\bar{E}_j\right), \tag{12}$$

where $\bar{E}_j$ is the aggregated PSVNHFN evaluated using the PSVNHFA operator in **(i)**, for $j = 1, \ldots, m$.

The computed criteria weights are then input into PSVNH–DLBD to benchmark the HMDTML models. In the subsequent section, this step is described in detail. The pseudocode of the PSVNH–FWZIC

method is shown in Algorithm 1.

Algorithm 1: **PSVNH-FWZIC method**

Step 1: Define the evaluation criteria.
Identify $C[j]$
Step 2: SEJ:
Define $Exp[i]$ //Identify the experts involved in the evaluation process.
Define EF, Imp //Identify the evaluation form.
$l \leftarrow$length (Exp)
For i =1: l **do**
  if $Exp(i)$ is true then
    $Exp(i) \leftarrow EF\ (i)$
  endif
End for
Step 3: Constructing EDM:
Initialize $EDM[i,j] \leftarrow Exp \uplus C$
$m \leftarrow$ length (C)
For i =1: l **do**
  For j =1:m **do**
    EDM $[i,j] \leftarrow$ Imp $(Exp[i]|C(j))$
  End for
End for
Step 4: **Apply PSVNH:**
For i =1: l **do**
  For j =1: m **do**
    $\tilde{E}[i,j] \leftarrow EDM[i,j]$ // Fuzzification using Table 2.
  End for
End for
Step 5: Compute criteria weight:
Step 5.1: Aggregate the criteria fuzzy values:
For j =1: m **do**
  $\bar{E}_j \leftarrow PSVNHFA\left(\tilde{E}[1,j], \tilde{E}[2,j], \dots, \tilde{E}[l,j]\right)$ //Aggregation using (10)
End for
Step 5.2: compute the score values of each criterion.
For j =1: m **do**
  $S[j] = s\left(\bar{E}_j\right)$ //Score calculation using (11)
End for
Step 5.3: Rescale the ultimate weight values.
For j =1: m **do**
  $w[j] = S[j] / \sum_{j=1}^{m} S[j]$ //Rescaling
End for

3.2.2 PSVNH–DLBD Benchmarking Method Formulation

In this section, a novel DLBD benchmarking method is proposed. The DLBD method is based on defining the length of scale using the maximum and minimum values of each criterion. The lower and upper boundaries for each scale are determined on the basis of the defined length of scale, and a dynamic

localisation decision matrix is formulated. Then, DLBD is extended under an FS environment (PSVNHF is used in this study) to generate a fuzzy dynamic localisation decision matrix. Finally, the alternatives are ranked to identify the best and worst alternatives. The proposed method could objectively benchmark the alternatives. The steps of PSVNH–DLBD formulation and development are as follows:

Step 1: Dynamic Localisation Decision Matrix Construction

(i) The initial decision matrix is formulated, as shown in Eq. (13). In this study, a uniform dynamic decision matrix of HMDTML models is constructed in Section 3.1 and employed in this section to create a dynamic localisation decision matrix. The decision matrix for n alternatives and m criteria has the following form:

$$\begin{matrix} & C_1 & C_2 & \cdots & C_m \\ A_1 \\ A_2 \\ \vdots \\ A_n \end{matrix} \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix}, \tag{13}$$

where x_{ij} is the i-th alternative (A_i) evaluation based on the j-th criterion (C_j).

(ii) The length of scale (L_j) is computed for each criterion in this step. In this context, each criterion's maximum value is subtracted from its minimum value, and the resulting value is divided by the number of scales (n_s), as shown in Eq. (14). n_s is set to five in the experiment based on the five-point Likert scale.

$$L_j = \frac{\max\limits_{i=1,\dots,n} x_{ij} - \min\limits_{i=1,\dots,n} x_{ij}}{n_s} \tag{14}$$

(iii) The lower bound (Lb) for each scale is computed, recursively. Under the cost-type criteria, the minimum value under a given criterion represents Lb of the positive ideal solution ($k = 1$). Meanwhile, Lb of the second scale ($k = 2$) is determined by summing the Lb value of the preceding scale and the L_j value of that criterion. The Lb values of the subsequent scale ($k = 3, 4, \dots, n_s$) are similarly determined, as shown in Eq. (15). The cost criterion in this study is log loss. If criterion C_j is cost-type, then the Lb values are

$$\begin{cases} Lb_1 = \min\limits_{i=1,\dots,n} x_{ij} \\ Lb_k = Lb_{k-1} + L_j, \text{for } k = 2, \dots, n_s \end{cases} \tag{15}$$

The Ub for each scale is computed. Under the cost-type criteria, the Ub of the positive ideal solution ($k = 1$) is determined by subtracting the ϵ value from the Lb of the second scale ($k = 2$). Meanwhile, the Lb of the second scale ($k = 2$) is determined by summing the Ub value of the preceding scale and the L_j value of that criterion and subtracting the ϵ value from the resulting value. The Ub values of the subsequent scale ($k = 3, 4, \dots, n_s$) are similarly determined, as shown in Eq. (16). If criterion C_j is cost-type, then

$$\begin{cases} Ub_1 = Lb_2 - \epsilon \\ Ub_k = Ub_{k-1} + L_j + \epsilon, \text{for } k = 2, \dots, n_s \end{cases} \quad (16)$$

where ϵ is a small positive number which separates the scales from each other.

(iv) Under the benefit-type criteria, the maximum value under a given criterion represents the Ub of the positive ideal solution ($k = 1$). Meanwhile, the Ub of the second scale ($k = 2$) is determined by subtracting the L_j value of that criterion from the Ub value of the preceding scale. The Ub values of the subsequent scale ($k = 3, 4, \dots, n_s$) are similarly determined, as shown in Eq. (17). If criterion C_j is benefit-type

$$\begin{cases} Ub_1 = \max\limits_{i=1,\dots,n} x_{ij} \\ Ub_k = Ub_{k-1} - L_j, \text{for } k = 2, \dots, n_s \end{cases} \quad (17)$$

Under the benefit-type criteria, Lb of the positive ideal solution ($k = 1$) is determined by subtracting the ϵ value (small value to separate scales) from the upper bound (Ub) of the second scale ($k = 2$). Meanwhile, Lb of the second scale ($k = 2$) is determined by subtracting the L_j value from the Lb value of the preceding scale and subtracting the resulting value from the ϵ value. The Lb values of the subsequent scale ($k = 3, 4, \dots, Ns$) are similarly determined, as shown in Eq. (18). The benefit criteria in this study are accuracy, specificity, precision, F1-score, recall and AUC. If criterion C_j is benefit-type, then the Lb values are

$$\begin{cases} Lb_1 = Ub_2 - \epsilon \\ Lb_k = Lb_{k-1} + L_j + \epsilon, \text{for } k = 2, \dots, n_s \end{cases} \quad (18)$$

where ϵ is a small positive number which separates the scales from each other.

(v) The initial decision matrix, which is a uniform dynamic decision matrix of HMDTML models, is transformed into a dynamic localisation decision matrix as follows:

$$D = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_n \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_m \end{array} \\ \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \dots & d_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ d_{n1} & d_{n2} & \cdots & d_{nm} \end{bmatrix} \end{array}, \quad (19)$$

$$\text{where } d_{ij} = \begin{cases} k & \text{if } \quad x_{ij} \in [Lb_k, Ub_k] \\ 0 & \text{otherwise} \end{cases}.$$

According to Eq. (19), each value in the decision matrix is replaced with a scale value ($k = 1, 2, \dots, n_s$) with respect to the upper and lower bounds of that scale.

Step 2: Fuzzy Dynamic Localisation Decision Matrix Creation

In this step, the scale values within the dynamic localisation decision matrix D, defined in Eq. (19),

are replaced with their associated PSVNHFNs to create a fuzzy dynamic localisation decision matrix, as shown in Eq. (20). Table 2 reports the scale values and their corresponding PSVNHFNs.

$$\tilde{d}_{ij} = \left(\mathrm{T}(d_{ij})|P^{T}(d_{ij}),\ \mathrm{I}(d_{ij})|P^{I}(d_{ij}),\ \mathrm{F}(d_{ij})|P^{F}(d_{ij})\right), \tag{20}$$

where $T(x), I(x)$ and $F(x)$ represent the hesitant degrees of possible truth, indeterminacy, and falsity for each scale, respectively.

Step 3: Benchmarking HMDTML Models

(i) The PSVNHFNs within the fuzzy dynamic localisation decision matrix are aggregated to find the fuzzy score value of each HMDTML model. Zhou *et al.* (Zhou et al., 2019) proposed an interval-valued neutrosophic Frank aggregation operator. The Probabilistic Single-valued Neutrosophic Hesitant Fuzzy Frank-weighed Averaging Aggregation (PSVNHFWA) operator is proposed in this work to find the fuzzy score value of each model based on their proposed aggregation operator, as shown in Eq. (21).

Let $N_j = \langle T_j \left| P^{T_j}, I_j \right| P^{I_j}, F_j \mid P^{F_j} \rangle$, for $j = 1, \ldots, r$, be PSVNHFNs, the PSVNHFWA operator aggregates r PSVNHFN into single PSVNHFN using a predefined weights $w = (w_1, w_2, \ldots., w_r)^T$ as follows:

$$\text{PSVNHFFWA}\ (N_1, N_2, \ldots, N_r) =$$

$$\left(\begin{array}{l} \bigcup_{(\alpha_j)_{j=1,\ldots,r} \in T_1 \times T_2 \times \ldots \times T_r} \left(1 - \log_\lambda\left(1 + \prod_{j=1}^{r} \left(\lambda^{1-\alpha_j} - 1\right)^{w_j}\right) \mid \log_\lambda\left(1 + \prod_{j=1}^{r} \left(\lambda^{P_{\alpha_j}^{T}} - 1\right)^{w_j}\right)\right), \\ \bigcup_{(\beta_j)_{j=1,\ldots,r} \in I_1 \times I_2 \times \ldots \times I_r} \left(\log_\lambda\left(1 + \prod_{j=1}^{r} \left(\lambda^{1-\beta_j} - 1\right)^{w_j}\right) \mid \log_\lambda\left(1 + \prod_{j=1}^{r} \left(\lambda^{P_{\beta_j}^{T}} - 1\right)^{w_j}\right)\right), \\ \bigcup_{(\gamma_j)_{j=1,\ldots,r} \in F_1 \times F_2 \times \ldots \times F_r} \left(\log_\lambda\left(1 + \prod_{j=1}^{r} \left(\lambda^{1-\gamma_j} - 1\right)^{w_j}\right) \mid \log_\lambda\left(1 + \prod_{j=1}^{r} \left(\lambda^{P_{\gamma_j}^{T}} - 1\right)^{w_j}\right)\right) \end{array}\right), \tag{21}$$

where $\lambda > 1$.

(ii) The fuzzy score values of all alternatives are defuzzied and turned into a numeric score value using the PSVNHFS score function shown in Eq. (11).

(iii) The models are reordered in a descending manner based on their score values, and the model with the highest score value is recommended (*rank=1*).

The pseudocode of the PSVNH–DLBD method is shown in Algorithm 2.

Step 2: Fuzzy Dynamic Localisation Decision Matrix Creation

In this step, the scale values within the dynamic localisation decision matrix are replaced with their associated PSVNHFNs to create a fuzzy dynamic localisation decision matrix, as shown in Eq. (20). Table 2 reports the scale values and their corresponding PSVNHFNs.

$$\widetilde{dlm}_{i,j} \approx \left(\tilde{t}(x), \tilde{\imath}(x), \tilde{f}(x)\right), \tag{20}$$

where $\tilde{t}(x), \tilde{\imath}(x)$ and $\tilde{f}(x)$ represent the hesitant degrees of possible truth, indeterminacy and falsity for each scale, respectively.

Step 3: Benchmarking HMDTML Models

(iv) The PSVNHFNs within the fuzzy dynamic localisation decision matrix are aggregated to find the fuzzy score value of each HMDTML model. Zhou *et al.* (Zhou et al., 2019) proposed an interval-valued neutrosophic Frank aggregation operator. The Probabilistic Single-valued Neutrosophic Hesitant Fuzzy Frank-weighed Averaging Aggregation (PSVNHFWA) operator is proposed in this work to find the fuzzy score value of each model based on their proposed aggregation operator, as shown in Eq. (21).

Let $\aleph_{\hat{\jmath}} = \left(\tilde{t}_{\hat{\jmath}} \middle| P^{\tilde{t}}{}_{\hat{\jmath}}, \tilde{\imath}_{\hat{\jmath}} \middle| P_{\hat{\jmath}}{}^{\tilde{\imath}}, \tilde{f}_{\hat{\jmath}} \middle| P_{\hat{\jmath}}{}^{\tilde{f}}\right)$ $(\hat{J} = 1,2, \ldots, r)$ be any group of PSVNHFNs and PSVNHFWA operator: PSVNHFN$^{\mathrm{r}}$ →PSVNHFN. Then, the PSVNHFFWA operator can be described as follows:

$$\text{PSVNHFFWA}\ (\aleph_1, \aleph_2, \ldots, \aleph_r) = \mathrm{W}_1 \cdot_F \aleph_1 \oplus_F \mathrm{W}_2 \cdot_F \aleph_2 \oplus_F \ldots . \oplus_F \mathrm{W}_r \cdot_F \aleph_r$$

where $\mathrm{W} = (\mathrm{W}_1, \mathrm{W}_2, \ldots . , \mathrm{W}_r)^T$ are the weights of $\aleph_{\hat{\jmath}} \in [0,1]$ with $\sum_{\hat{\jmath}=1}^{r} \mathrm{W}_{\hat{\jmath}} = 1$. Then, the aggregation result using PSVNHFFWA is as follows:

$$\text{PSVNHFFWA}\ (\aleph_1, \aleph_2, \ldots, \aleph_r) =$$

$$\left(\begin{array}{c} \bigcup_{\alpha_{\aleph_{\hat{\jmath}}} \in t_{\aleph_{\hat{\jmath}}}, \tilde{p}^{\alpha}{}_{\aleph_{\hat{\jmath}}} \in \tilde{p}^{t}{}_{\aleph_{\hat{\jmath}}}} 1 - \log_\lambda \left(1 + \prod_{\hat{\jmath}=1}^{r} \left(\lambda^{1-\alpha_{\aleph_{\hat{\jmath}}}} - 1\right)^{\mathrm{W}_{\hat{\jmath}}}\right) / \log_\lambda \left(1 + \prod_{\hat{\jmath}=1}^{r} \left(\lambda^{\tilde{p}^{\alpha}{}_{\aleph_{\hat{\jmath}}}} - 1\right)^{\mathrm{W}_{\hat{\jmath}}}\right), \\ \bigcup_{\beta_{\aleph_{\hat{\jmath}}} \in i_{\aleph_{\hat{\jmath}}}, \tilde{p}^{\beta}{}_{\aleph_{\hat{\jmath}}} \in \tilde{p}^{i}{}_{\aleph_{\hat{\jmath}}}} \log_\lambda \left(1 + \prod_{\hat{\jmath}=1}^{r} \left(\lambda^{\beta_{\aleph_{\hat{\jmath}}}} - 1\right)^{\mathrm{W}_{\hat{\jmath}}}\right) / \log_\lambda \left(1 + \prod_{\hat{\jmath}=1}^{r} \left(\lambda^{\tilde{p}^{B}{}_{\aleph_{\hat{\jmath}}}} - 1\right)^{\mathrm{W}_{\hat{\jmath}}}\right), \\ \bigcup_{\gamma_{\aleph_{\hat{\jmath}}} \in f_{\aleph_{\hat{\jmath}}}, \tilde{p}^{\gamma}{}_{\aleph_{\hat{\jmath}}} \in \tilde{p}^{f}{}_{\aleph_{\hat{\jmath}}}} \log_\lambda \left(1 + \prod_{\hat{\jmath}=1}^{r} \left(\lambda^{\gamma_{\aleph_{\hat{\jmath}}}} - 1\right)^{\mathrm{W}_{\hat{\jmath}}}\right) / \log_\lambda \left(1 + \prod_{\hat{\jmath}=1}^{r} \left(\lambda^{\tilde{p}^{\gamma}{}_{\aleph_{\hat{\jmath}}}} - 1\right)^{\mathrm{W}_{\hat{\jmath}}}\right) \end{array}\right), \tag{21}$$

where $\lambda > 1$

(v) The fuzzy score values of all alternatives are defuzzied and turned into a numeric score value using the PSVNHFS score function shown in Eq. (11).

(vi) The models are reordered in a descending manner based on their score values, and the model with the highest score value is recommended (*rank=1*).

The pseudocode of the PSVNH–DLBD method is shown in Algorithm 2.

Algorithm 2: PSVNH–DLBD Benchmarking Method	
Input: Models, performance evaluation criteria and criteria weights, numbers of scale (n_s), epsilon value (ϵ) **Output:** Benchmark Models	
1: **Step 1: Decision Matrix Formulation:**	
2: **Identify** $C[j]$	//Criteria
3: **Identify** A $[i]$	//Alternatives
4: n ← length (C)	
5: m ← length (A)	

$X[i,j] \leftarrow A \cap C$ //Decision Matrix

Step 1.1: Input criteria Categorisation (Benefit [Be] and Cost [Co])

Step 1.2: Calculate the length of scale for each criterion.

Identify n_s //Number of scales

Identify ϵ //Small number for separation

For $j = 1:m$

  $L[j] = (\max_{i=1,\dots,n} X[i,j] - \min_{i=1,\dots,n} X[i,j])/n_s$

End for

Step 1.3: Calculate the Lb and Ub for each scale.

For $j = 1:m$

  $Lb[1] = \min_{i=1,\dots,n} X[i,j]$

  For $k = 2:n_s$

    $Lb[k] = Lb[k-1] + L[j]$

  End for

  $Ub[1] = Lb[2] - \epsilon$

  For $k = 2:n_s$

    $Ub[k] = Ub[k-1] + L[j] + \epsilon$

  End for

  If $C[j]$ is benefit type

    $Lb \leftarrow \mathbf{flip}(Lb)$

    $Ub \leftarrow \mathbf{flip}(Ub)$

  End If

Step 1.4: Transform the decision matrix into dynamic localisation decision matrix.

  For $i = 1:n$

    For $k = 1:n_s$

      If $X[i,j] \in [Lb[k], Ub[k]]$

        $D[i,j] = k$

      End if

    End for

  End for

End For

Step 2: Transform dynamic localisation decision matrix into Fuzzy Dynamic Localisation Decision Matrix Creation

For $i = 1:n$

  For $j = 1:m$

    $\tilde{D}[i,j] = \left(T(D[i,j])|P^T(D[i,j]),\ I(D[i,j])|P^I(D[i,j]),\ F(D[i,j])|P^F(D[i,j])\right)$

  End for

End for

Step 3: Compute the scores of the alternatives:

Identify λ

Step 3.1: Find the aggregation value for each alternative.

For $i = 1:n$

  $Agg[i] = \text{PSVNHFFWA}\left(\tilde{D}[i,1], \dots, \tilde{D}[i,m]\right)$ // Aggregation using (21) and

  //weight coefficients imported from Algorithm 1

End for

Step 3.2: Find the score value for each alternative.

For $i = 1{:}n$
    $Score[i] = s(Agg[i])$ //Score calculation using (11)
End for
Step 3.3: Benchmark HMDTML Models.

4 Results and Discussion

4.1 Results of Uniform Dynamic Decision Matrix

This section reports the results of the uniform dynamic decision matrix. Four pretrained CNN models (Painters, VGG19, DeepLoc and SqueezeNet) are integrated with 12 machine learning algorithms (Naive Bayes, Neural Network, SVM [linear], RBF, kNN, SGD, Logistic Regression, Random Forest, Decision Tree, AdaBoost, CN2 rule inducer and SVM [Polynomial]). The resulting forty-eight HMDTML models are used to detect COVID-19 CXR images. The results of the HMDTML models are evaluated using seven performance evaluation criteria (AUC, accuracy, F1-score, precision, recall, time and specificity). Table A.1 (Appendix A) presents an evaluation of the 48 HMDTML models based on the seven evaluation criteria. The resulting uniform dynamic decision matrix is fed to the PSVNH–DLBD benchmarking method to rank the HMDTML model alternatives.

4.2 Weight estimation of HMDTML Model Criteria

The performance evaluation criteria of the HMDTML models, which have been defined and identified in Section 3.1 for benchmarking the HMDTML model, have been subjected to weighting through the PSVNH-FWZIC method, resulting in achievement of outcomes. One of the main advantages of utilising PSVNH-FWZIC is the consistencies in the calculated weights. As described in Section 3.2.1, the PSVNH–FWZIC method has five steps. In the first step, the set of evaluation criteria was examined and investigated. In the second step, three experts were recruited to evaluate all the evaluation criteria using five linguistic expressions and an evaluation form. Table 3 depicts the outcome of this particular step, which is the SEJ. In the third step, the EDM has been built by replacing with a numerical scale, as given in Table 3. In the fourth step, the PSVNHFS–EDM was constructed by replacing a numeric scale with PSVNHFNs. In the fifth and last step, the final weights have been calculated by aggregating and defuzzing the PSVNHFNs of each criterion within the PSVNHFS–EDMs, as given in Table 3.

Table 3
Criterion weighting results of HMDTML models.

Criteria/Experts	SEJ			EDM			Final Weights
	Expert 1	Expert 2	Expert 3	Expert 1	Expert 2	Expert 3	
C_1	VI	VI	VI	1	1	1	0.1555
C_2	I	I	VI	2	2	1	0.1455
C_3	I	I	I	2	2	2	0.1354

C_4	LI	Av	Av	4	3	3	0.1265
C_5	Av	LI	I	3	4	2	0.1306
C_6	VI	VI	VI	1	1	1	0.1555
C_7	I	VI	VI	2	1	1	0.1510

As presented in Table 3, AUC (C_1) and Time (C_6) were the criteria that achieved the highest weighting value, specifically 0.1555. This was closely followed by the Specificity criterion (C_7) which was assigned a weighting of 0.1510. Further, the weighting value attributed to the Classification Accuracy criterion (C_2) was 0.1455. Subsequently, the F1-score (C_3) and Recall (C_5) criteria received weighting values of 0.1354 and 0.1306 respectively. The Precision criterion (C_4) was assigned the lowest weighting value, at 0.1265. As previously detailed, these generated weights, in conjunction with the uniform dynamic decision matrix, are utilized in ranking methodologies to benchmark the HMDTML models.

4.3 HMDTML Model Benchmarking Results

This section reports the benchmarking results of the HMDTML models using the PSVNH–DLBD method detailed in Section 3.2.2. The PSVNH–DLBD method consisted of three main steps: dynamic localisation decision matrix construction, fuzzy dynamic localisation decision matrix creation and benchmarking HMDTML models. In the first step, the uniform dynamic decision matrix of the HMDTML models formulated in Section 3.1 is employed to create a dynamic localisation decision matrix. Subsequently, the length of scale (L_j) was calculated for each criterion, as given in Table 4. Then, the lower bound (Lb) and upper bound (Ub) were calculated for each scale based on the maximum and minimum values of each criterion and the calculated L_j, as given in Table 4.

Table 4
Resulting $L_j, min, max, Lb \ and \ Ub$ values with their associated scales.

Criteria	L_j	min	max	Lb	Ub	Scale
				0.9857	0.9998	1
				0.9716	0.9857	2
C_1	0.0141	0.9292	0.9998	0.9575	0.9716	3
				0.9434	0.9575	4
				0.9292	0.9434	5
				0.9272	0.9948	1
				0.8596	0.9272	2
C_2	0.0676	0.6570	0.9948	0.7921	0.8596	3
				0.7245	0.7921	4
				0.6570	0.7245	5
				0.9182	0.9948	1
C_3	0.0765	0.6121	0.9948	0.8417	0.9182	2
				0.7652	0.8417	3

				0.6886	0.7652	4
				0.6121	0.6886	5
				0.9542	0.9948	1
				0.9137	0.9542	2
C_4	0.0405	0.7920	0.9948	0.8731	0.9137	3
				0.8326	0.8731	4
				0.7920	0.8326	5
				0.9272	0.9948	1
				0.8596	0.9272	2
C_5	0.0676	0.6570	0.9948	0.7921	0.8596	3
				0.7245	0.7921	4
				0.6570	0.7245	5
				0.0187	0.5821	1
				0.5821	1.1454	2
C_6	0.5634	0.0187	2.8356	1.1454	1.7088	3
				1.7088	2.2722	4
				2.2722	2.8356	5
				0.9272	0.9948	1
				0.8596	0.9272	2
C_7	0.0676	0.6570	0.9948	0.7921	0.8596	3
				0.7245	0.7921	4
				0.6570	0.7245	5

Thereafter, the uniform dynamic decision matrix of the HMDTML models was transformed into a dynamic localisation decision matrix, as presented in Table A.2 (Appendix A). In the second step, the fuzzy dynamic localisation decision matrix was created by replacing the scale values within the dynamic localisation decision matrix with their corresponding PSVNHFNs. In the third step, these PSVNHFNs were aggregated using the PSVNHFWA operator and defuzzied, and the final score values were achieved, as given in Table 5. Finally, the alternatives were ranked based on their score values, with the alternative with the highest score value ranked first.

Table 5
Benchmarking results of HMDTML models ($\boldsymbol{\lambda = 2}$).

HMDTML Models		Scores	Ranks	HMDTML Models		Scores	Ranks
M1	Painters-Naive Bayes	0.2438	6	**M25**	DeepLoc-Naive Bayes	0.2474	2
M2	Painters-Neural Network	0.2386	19	**M26**	DeepLoc-Neural Network	0.2386	19
M3	Painters-SVM (linear)	0.2386	19	**M27**	DeepLoc-SVM (linear)	0.2386	19
M4	Painters-RBF	0.2413	14	**M28**	DeepLoc-RBF	0.2438	6
M5	Painters-kNN	0.2386	19	**M29**	DeepLoc-kNN	0.2386	19
M6	Painters-SGD	0.2386	19	**M30**	DeepLoc-SGD	0.2386	19
M7	Painters-Logistic Regression	0.2386	19	**M31**	DeepLoc-Logistic Regression	0.2386	19
M8	Painters-Random Forest	0.2432	10	**M32**	DeepLoc-Random Forest	0.2432	10

M9	Painters-Decision Tree	0.2456	4	**M33**	DeepLoc-Decision Tree	0.2352	47
M10	Painters-AdaBoost	0.2386	19	**M34**	DeepLoc-AdaBoost	0.2474	3
M11	Painters-CN2 rule inducer	0.2386	19	**M35**	DeepLoc-CN2 rule inducer	0.2386	19
M12	Painters-SVM (Polynomial)	0.2438	6	**M36**	DeepLoc-SVM (Polynomial)	0.2400	17
M13	VGG19-Naive Bayes	0.2386	19	**M37**	SqueezeNet-Naive Bayes	0.2386	19
M14	VGG19-Neural Network	0.2386	19	**M38**	SqueezeNet-Neural Network	0.2386	19
M15	VGG19-SVM (linear)	0.2386	19	**M39**	SqueezeNet-SVM (linear)	0.2055	48
M16	VGG19-RBF	0.2432	10	**M40**	SqueezeNet-RBF	0.2432	13
M17	VGG19-kNN	0.2392	18	**M41**	SqueezeNet-kNN	0.2432	9
M18	VGG19-SGD	0.2386	19	**M42**	SqueezeNet-SGD	0.2386	19
M19	VGG19-Logistic Regression	0.2386	19	**M43**	SqueezeNet-Logistic Regression	0.2386	19
M20	VGG19-Random Forest	0.2413	14	**M44**	SqueezeNet-Random Forest	0.2413	14
M21	VGG19-Decision Tree	0.2386	19	**M45**	SqueezeNet-Decision Tree	0.2386	19
M22	VGG19-AdaBoost	0.2386	19	**M46**	SqueezeNet-AdaBoost	0.2386	19
M23	VGG19-CN2 rule inducer	0.2386	19	**M47**	SqueezeNet-CN2 rule inducer	0.2386	19
M24	VGG19-SVM (Polynomial)	0.2509	1	**M48**	SqueezeNet-SVM (Polynomial)	0.2453	5

Table 5 presents the score and rank values of each alternative when $\lambda = 2$. Table A.3 (Appendix A) presents the benchmarking result of the HMDTML models when $\lambda = 3, \lambda = 4, \lambda = 5, \lambda = 6, \lambda = 7, \lambda = 8, \lambda = 9$ and $\lambda = 10$. According to Table 5 and Table A.3, Model M24 (VGG19-SVM [Polynomial]) earned the highest rank (*rank=1*) when $\lambda = 2, 3, 4, 5$ and 6 with score values of 0.2509, 0.2523, 0.2532, 0.2538 and 0.2543, respectively, and it dropped to the second rank (rank=2) when $\lambda = 7, 8, 9$ and 10 with score values of 0.2548, 0.2551, 0.2554 and 0.2557, respectively. Model M25 (DeepLoc-Naive Bayes) received the second rank when $\lambda = 2, 3, 4, 5$ and 6 with score values of 0.2474, 0.2499, 0.2516, 0.2528 and 0.2538, respectively; however, it raised to the first rank when $\lambda = 7, 8, 9, 10$ with score values of 0.2546, 0.2553, 0.2559, and 0.2564, respectively. Model M34 (DeepLoc-AdaBoost) obtained the third rank when $\lambda = 2, 3, 4, 5$ and 6 had score values of 0.2474, 0.2495, 0.2510, 0.2520 and 0.2529; however, it dropped to the fourth rank when $\lambda = 7$ had a score value of 0.2535 and to the fifth rank when $\lambda = 8, 9$ and 10 had score values of 0.2541, 0.2546 and 0.2550. Model M39 (SqueezeNet-SVM [linear]) had the lowest rank (rank=48) across all λ values, with score values of 0.2055, 0.2081, 0.2099, 0.2112, 0.2122, 0.2131, 0.2138, 0.2144 and 0.2149. The ranking results of the remaining models were distributed between the highest and lowest models. Overall, the λ values were set to 2, 3, 4, 5, 6, 7, 8, 9 and 10 to explore the changes in the final ranking results.

5 Evaluation and Validation

5.1 Sensitivity analysis

The proposed PSVNH-FWZIC and PSVNH-DLBD methods are subjected to sensitivity analysis test to evaluate their robustness. This test has been employed by various researchers to verify their innovations

(Qumar, et al., 2022). Upon calculation of the weight value for each criterion through the employment of PSVNH-FWZIC (see Table 3), the criterion deemed as the 'most important' was identified for the purpose of conducting sensitivity analysis. The objective of the test is to assess the impact of the "most significant criterion" on the benchmarking outcomes of the proposed method. According to Pamucar et al. (2020) identified the proportionality of the criteria's weights during sensitivity analysis using Eq. (22) and determined the elasticity coefficient (e_c). The e_c value was used the relative modification of the remaining weight coefficients as a result to the changes in the weight coefficient of the important criterion.

$$w_c = w_c^o - \delta\, e_c, \tag{22}$$

where $e_c = 1/(1 - w_s^0)$ and w_s^o is the weight of the most important criterion. The criteria weights obtained through the PSVNH-FWZIC method are represented by w_c^o, while δ is a parameter controlling the size of the change in the weight.

The present study has identified AUC (C_1) and time (C_6) as the most important criteria, owing to their relatively higher weight value ($w_{AUC} = w_{time} = 0.1555$). Thereafter, the e_c values corresponding to the criteria are established, as given in Table 6.

Table 6
Computed e_c for changing weights.

Criteria	C_1	C_2	C_3	C_4	C_5	C_6	C_7
e_c values	0.1842	0.1723	0.1603	0.1498	0.1547	0.1842	0.1788

The δ values signify the extent of changes made to a set of criteria weight values, and it depends on the e_c values. The values of δ could be either positive or negative, signifying an increase or decrease in relative importance, respectively. The bounds of δ are defined by the maximum adjustments in the weight of the most important criterion in both the positive and negative aspects. Accordingly, the δ limit values of C_1 and C_6 are determined by the interval $[\delta_{lb}, \delta_{ub}] = [-0.1555, 0.8445]$, where $\delta_{lb} = -\max_{j=1,\dots,m} w_j$ and $\delta_{ub} = 1 - \max_{j=1,\dots,m} w_j$. The sensitivity analysis scenarios were identified based on the determined limit values. Accordingly, $-0.1555 \le \delta \le 0.8445$ was split into nine scenarios for $(\delta_i)_{i=1,\dots,10}$, where

$$\begin{cases} \delta_1 = \delta_{lb} \\ \delta_i = \delta_{i-1} + \frac{\delta_{ub} - \delta_{lb}}{9}, \quad \text{for } 2 \le i \le 9\,. \\ \delta_{10} = \delta_{ub} \end{cases}$$

Table 7 highlights the criteria weight list utilised in the sensitivity analysis of the nine scenarios. It is necessary for the weight values to satisfy the condition of the summation of w_c being equal to 1, and hence the weight of the most sensitive criteria is updated as $w_s = 1 - \sum_{c \ne s} w_c$. The table presents the new weights when the most important criterion AUC is set to zero in the first scenario and when the most important criterion time is set to zero in the first scenario.

Table 7
New criterion weights for sensitivity analysis.

Criteria/scenarios		PSVNH–FWZIC	S1	S2	S3	S4	S5	S6	S7	S8	S9
The most important criterion is AUC.	C_1	0.1555	0.0000	0.1250	0.2500	0.3750	0.5000	0.6250	0.7500	0.8750	0.9999
	C_2	0.1455	0.1723	0.1507	0.1292	0.1077	0.0861	0.0646	0.0431	0.0215	0.0000
	C_3	0.1354	0.1603	0.1403	0.1202	0.1002	0.0802	0.0601	0.0401	0.0200	0.0000
	C_4	0.1265	0.1498	0.1311	0.1124	0.0936	0.0749	0.0562	0.0375	0.0187	0.0000
	C_5	0.1306	0.1547	0.1353	0.1160	0.0967	0.0773	0.0580	0.0387	0.0193	0.0000
	C_6	0.1555	0.1842	0.1611	0.1381	0.1151	0.0921	0.0691	0.0460	0.0230	0.0000
	C_7	0.1510	0.1788	0.1564	0.1341	0.1117	0.0894	0.0670	0.0447	0.0223	0.0000
The most important criterion is time	C_1	0.1555	0.1842	0.1611	0.1381	0.1151	0.0921	0.0691	0.0460	0.0230	0.0000
	C_2	0.1455	0.1723	0.1507	0.1292	0.1077	0.0861	0.0646	0.0431	0.0215	0.0000
	C_3	0.1354	0.1603	0.1403	0.1202	0.1002	0.0802	0.0601	0.0401	0.0200	0.0000
	C_4	0.1265	0.1498	0.1311	0.1124	0.0936	0.0749	0.0562	0.0375	0.0187	0.0000
	C_5	0.1306	0.1547	0.1353	0.1160	0.0967	0.0773	0.0580	0.0387	0.0193	0.0000
	C_6	0.1555	0.0000	0.1250	0.2500	0.3750	0.5000	0.6250	0.7500	0.8750	0.9999
	C_7	0.1510	0.1788	0.1564	0.1341	0.1117	0.0894	0.0670	0.0447	0.0223	0.0000

Fig.4 depicts the influence of the generated weight values (see Table 7) on the ranking orders of the initial 10 models ($\lambda = 2$). The study conducted a comparison of the initial 10 models, namely M24, M25, M34, M9, M48, M1, M12, M28, M41, and M8, under the condition where the weight value of the AUC criterion is set to zero in the first scenario. The results indicate that Model M24 consistently maintained the top rank in five scenarios (S1-S5), while it was ranked second in S6, 38 in S7, and 42 in S8 and S9, as illustrated in Fig.4(a). Model M25 attained the second position in three out of five scenarios (S3-S5), whereas it was dropped in the remaining scenarios. The Model M34 was ranked third in four scenarios, namely S1 and S3-S5, whereas its ranking varied in the remaining scenarios. In four out of the total scenarios (S2-S5), Models M9 and M48 were placed in the fourth and fifth positions, respectively. However, these models were dropped in the remaining scenarios. The rankings of Models M1, M12, M28, M41 and M8 were completely changed in all scenarios. Figs. B1–B8 (Appendix B) present the influence of the new weight values (see Table 7) on the ranking order of the first 10 models for λ values of 3–10.

According to the comparison of the first 10 models when the weight value of the time criterion is set to zero in the first scenario, Model M24 maintained the first rank in all scenarios, as shown in Fig.4(b). The M25 model achieved the second position in two scenarios, namely S3 and S4, whereas it was dropped in the remaining scenarios. The results indicate that Model M34 achieved the third rank in scenarios S3, S4, and S9, whereas it attained the second rank in the remaining scenarios. In scenarios S3 and S4, Model M9 attained the fourth position, whereas it was dropped in the remaining scenarios. In one scenario, the rankings of Models M48 and M41 remained consistent at fifth and ninth place, respectively, whereas in the remaining scenarios, their rankings were altered. The models M1, M12, and M28 attained the sixth position in three

distinct scenarios (S2-S4), whereas they were dropped in the remaining scenarios. According to the results, Model M8 exhibited a ranking of tenth place in scenarios S2, S3, and S5, whereas it demonstrated an increase in ranking in the remaining scenarios. Fig. B9–B16 (Appendix B) present the influence of the new weight values (see Table 7) on the ranking order of the first 10 models for λ values of 3–10.

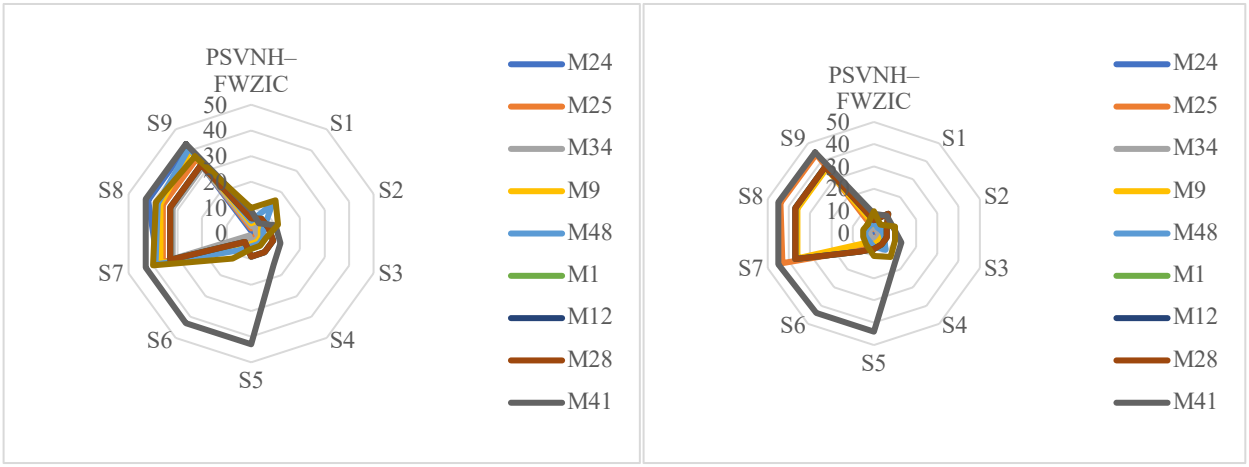


Fig.4. Sensitivity analysis results ($\boldsymbol{\lambda = 2}$) with the weight of the most important criteria (a) AUC and (b) time set to zero in the first scenario.

The findings, as illustrated in Fig.4, demonstrate that the ranking order of the most models are modified by the application of the generated weight values in the nine scenarios. Therefore, it can be inferred that the suggested method is responsive to variations in weight values.

The Spearman's rank correlation coefficient (SCC) is utilized to accurately assess the correlation between the new and initial outcomes, given that the proposed methodology has demonstrated sensitivity to variations in the criteria weights. Fig. 5 depicts the representation of the SCC values for the nine scenarios across all λ values.

First, an evaluation was performed on the SCC values, with the weight value of the AUC criterion being set to zero in the first scenario, as illustrated in Fig. 5(a). In the initial scenario, strong and positive correlations were observed between the initial ranking results and the new one when λ values were set to 2, 4, and 5. Moderate and positive correlations were observed when λ values were set to 3, 6, 7, 8, 9, and 10. The correlations exhibited a strong and positive association in the second and third scenarios, consistently across all λ values. The fourth and fifth scenarios exhibited moderate and positive correlations at λ=2, while the remaining λ values demonstrated strong and positive correlations. The sixth scenario exhibited moderate and positive correlations at λ=2 and λ=3, while the remaining λ values showed weak and positive correlations. The seventh, eighth, and ninth scenarios exhibited negative and weak correlations at λ=2,

while displaying weak and positive correlations at all other λ values. The study found that the mean of the SCC values for all scenarios was 0.6 when λ values were between 3 and 10, indicating a moderate positive correlation. However, when λ was 2, the mean of SCC values was 0.5, indicating a weak positive correlation.

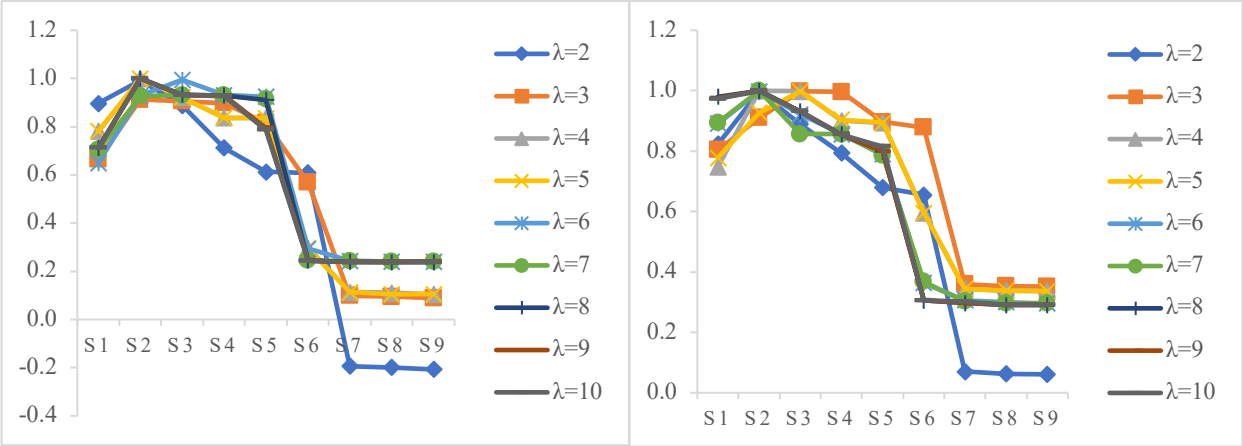


Fig. 5. SCC results ($\boldsymbol{\lambda = 2 - \lambda = 10}$) with the weight of the most important criteria (a) AUC and (b) time set to zero in the first scenario.

Second, an evaluation was performed on the SCC values, with the weight value of the time criterion is set to zero in the first scenario, as illustrated in Fig. 5(b). The initial ranking results exhibited strong and positive correlations with the new ranking results across all λ values in the first scenario, with the exception of λ=4, which demonstrated a moderate and positive correlation. Strong and positive correlations were observed across all λ values in the second, third, and fourth scenarios. Across all λ values, except for λ=2, the correlations exhibited a strong and positive association in the fifth scenario. However, for λ=2, the correlation was moderate and positive. In the sixth scenario, a moderate and positive correlation was observed at λ values of 2, 4, and 5, a strong and positive correlation at a λ value of 3, and a weak and positive correlation at all other λ values. Across all λ values, the correlations in scenarios seven, eight, and nine were negative and weak. The study found that the mean of SCC values was 0.6 for λ values of 2, 6, 7, 8, 9, and 10, and 0.7 for λ values of 3, 4, and 5, indicating a moderate positive correlation.

5.2 Comparative analysis

This section includes two types of comparisons. The first comparison assessed the DLBD proposed with the FDOSM method (Salih et al., 2020). This evaluation was based on eleven theoretical points, as detailed in Section 5.2.1. The second comparison examined the present study in comparison to a previous study conducted by (M. Mohammed et al., 2020). This comparison focused on the application and theory aspects, as detailed in Section 5.2.2.

5.2.1 Comparative analysis based on theoretical points

This section compares the proposed DLBD against FDOSM (Salih et al., 2020) based on the eleven theoretical points. The comparison points presented in Table 8 are summarised as follows:

1st point: *Ranking method features:*

The first subpoint refers to the suitability of the ranking method (DLBD or FDOSM) for subjective or objective evaluation with a dynamic change in value.

The second subpoint refers to the ideology of the ranking method.

The 3rd subpoint refers to the nature of the generated matrix as the basis for evaluation.

2nd point: *Processing time* (Salih et al., 2020) refers to the time required to perform localisations or comparisons.

3rd point: *Unnatural comparisons* (Rozi et al., 2022; Salih et al., 2020) refer to the ability of the ranking method to compare two uncorrelated criteria based on a scale (e.g. one to nine) to indicate the number of times this specific criterion is better than the other criteria.

4th point: *Inconsistency* (Salih et al., 2020) refers to the ability of the ranking method to solve this problem, which is the primary issue with subjective methods.

5th point: *Distance measurement problem* (Salih et al., 2020) refers to the ability of the ranking method to solve this problem.

6th point: *Normalisation* (Rozi et al., 2022; Salih et al., 2020) refers to the requirement of the ranking method to normalise the values of the decision matrix, which is an issue owing to the availability of several normalisation formulas that affect the final ranking results.

7th point*: Implicitly weighting criterion values* (Salih et al., 2020) refers to the ability of the ranking method to implicitly weight certain criterion values.

8th point: *Missing information* (Rozi et al., 2022; Salih et al., 2020) refers to the ability of the ranking method to solve this problem.

9th point: *Immeasurable values* (Salih et al., 2020) refer to the efficiency of the ranking method to handle values that cannot be quantified, such as binomial criteria (e.g. yes/no), polynomial criteria (e.g. colours), textual data (e.g. brand names), and categorical criteria (e.g. interval values, ranges).

10th point: *Vagueness and ambiguous information* (Salih et al., 2020) refers to the ability of the ranking method to solve this problem.

11th point: *Explicit weighting* (Rozi et al., 2022; Salih et al., 2020) refers to the ability of the ranking method to explicitly estimate the weight of the evaluation criteria.

Table 8
Comparison points between DLBD and FDOSM.

Comparison points	DLBD	FDOSM
Ranking method features	Objective evaluation with a dynamic change of values	Subjective evaluation
	Employing upper and lower boundaries of each criterion.	Employing ideal solutions of each criterion
	Dynamic localisation	Opinion
Processing time	Minimise the processing time required for localisation	Minimise the processing time required for comparisons
Unnatural comparisons	Not applicable	Solved
Inconsistency	Not applicable	Solved
Distance measurement	Solved	Solved
Normalisation	Solved	Solved
Implicitly weighting criterion values	Solved based on the localised values within their boundaries	Solved based on experts' opinion of each value
Missing information	Unconsidered for objective values	Solved for subjective values
Immeasurable value	Not applicable	Solved
Ambiguous and vagueness Information	Solved with applied FSs	Solved
Explicit weighting	Require an external method	Require an external method

According to Table 8, the DLBD method was compared with FDOSM based on eleven theoretical points. The first point (ranking method features) compared the two ranking methods according to three subpoints. The first subpoint demonstrated that the DLBD method is most suitable to evaluate objective data with a dynamic change in values. Meanwhile, FDOSM is most suitable for evaluating subjective data. The second subpoint compared the ideology of the two methods. DLBD utilised the upper and lower boundaries of each criterion, whereas FDOSM used the positive and negative ideal solutions for each criterion. The third subpoint concerned the nature of the created matrix as the evaluation's foundation. DLBD generates a dynamic localisation decision matrix as a basis for evaluating alternatives. Meanwhile, FDOSM generates an opinion matrix for the same purpose. The second point compared the processing time requirements of the two methods. DLBD reduces the processing time required for localisation, while FDOSM minimises the processing time required for comparison. The third and fourth points evaluated the ranking method's ability to assess two uncorrelated criteria and overcome an inconsistency problem, respectively. These issues are raised with a subjective evaluation. Accordingly, FDOSM successfully handled both issues, but it is not applicable in objective evaluation. The fifth and sixth points referred to the ability of the ranking method to address distance measurement and normalisation problems,

respectively. FDOSM used positive and negative ideal solutions to successfully solve these issues by eliminating the use of the normalisation method and distance measurement, while DLBD used lower and upper boundaries for the same purpose. The seventh point referred to the ability of the ranking method to implicitly weight certain criterion values. DLBD implicitly weighted certain criterion values based on localising values within their boundaries. Meanwhile, FDOSM implicitly weighted certain criterion values based on experts' opinions of each value. The eighth point referred to the ability of the ranking method to solve missing information within a decision matrix. Salih et al. (Salih et al., 2020) stated that FDOSM handled this issue; nevertheless, they provided no instances of this issue. In addition, DLBD did not consider the missing information of objective values. Therefore, FDOSM and DLBD should consider this issue in the future. The ninth point referred to the ability of the ranking method to deal with immeasurable values. These types of values are experienced with subjective evaluation; hence, FDOSM successfully handled them, but it is not applicable in objective evaluation. The tenth point referred to the ability of the ranking method to solve vagueness and ambiguous information. The DLBD method solved this problem by applying FSs (e.g. PSVNH). However, the ideology of FDOSM must apply FS to overcome this problem. The eleventh point referred to the ability of the ranking method to explicitly estimate the weight of the evaluation criteria. Both methods required an external weighting method to explicitly estimate the weight of the evaluation criteria.

Overall, the analysis of the comparisons revealed that the DLBD method and FDOSM solved most MCDM theoretical issues. However, the DLBD method is optimal for evaluating objective data with dynamic changes in values, whereas the FDOSM method is optimal for evaluating subjective data.

5.2.2 Comparative analysis in terms of application and theory

Many remarkable studies are available in the literature such as (Cai et al., 2023; Hu et al., 2021; Khishe et al., 2021; Saffari et al., 2022; Wang et al., 2022). However, they did not consider using MCDM methods to rank the proposed models based on the seven predefined matrices. Therefore, this section compares the proposed PSVNH-FWZIC and PSVNH-DLBD MCDM methods with the entropy and TOPSIS methods used in a previous study (M. Mohammed et al., 2020). The comparison is based on 18 points. The present comparison examines both the application and theory aspects of the evaluation and benchmarking of HMDTML alternatives in the context of MCDM. On the first hand, the application aspect focuses on comparing the MCDM issues associated with the evaluation and benchmarking of HMDTML alternatives. On the other hand, the theory aspect involves comparing the weighing and benchmarking methods employed in the current study with those used in (M. Mohammed et al., 2020). The application aspect encompasses three comparison points, while the theory aspect encompasses a total of 15 comparison points. Among these, seven points fall under the weighting perspective, while the remaining eight points fall under

the ranking perspective. These comparison points are detailed in Table 9.

Table 9
Comparison points between the compared studies.

Comparison Points			The current study	The study of (M. Mohammed et al., 2020)
Application Aspect	Several criteria (matrices) were considered.		√	√
	The criteria were evaluated according to their level of importance.		√	√
	The issue in data variations was addressed.		√	√
Theory Aspect	Weighting perspective	The issue of inconsistency regarding the weighting method is handled.	√	√
		The utilization of pairwise comparisons is not required by the weighting method.	√	√
		The weighting method is characterised by its simplicity in implementation and its lack of dependence on data analysis.	√	×
		The weighting method possesses inherent characteristics that enhance transparency.	√	×
		The weighting method provides a degree of adaptability in accommodating a wide range of perspectives and values.	√	×
		The issue concerning the vagueness, uncertainty and hesitancy that arise from data or expert preferences in the weighting method is effectively solved.	√	×
		The utilization of FSs in the weighting method has a broad range of options.	√	×
	Benchmarking perspective	The benchmarking method can assign weights to the evaluation attributes.	×	×
		The benchmarking method effectively manages dynamic changes by considering the consequences of boundary values.	√	×
		The benchmarking method can reduce the variance of outcomes across normalization methods.	√	×
		The benchmarking method decreases the processing time required to perform localization	√	×
		The benchmarking method uses upper and lower boundaries to overcome the distance measurement problem	√	×
		The benchmarking method can also automatically estimate the implicit weights for the set of values in each criterion by using scale values.	√	×
		The issue concerning the vagueness, uncertainty and hesitancy that arise from data or expert preferences in the benchmarking method is effectively solved.	√	×
		The utilization of FSs in the benchmarking method has a broad range of options.	√	×
Total score			94.4%	27.8%
Accumulative difference			5.6%	72.2%

The present study and a previous study (M. Mohammed et al., 2020) have met the initial three points (100%) in terms of application, as shown in Table 9. Both studies examined the inconsistency associated with weighting methods and did not involve pairwise comparisons.

However, the weighting method used in the present study differs from the one used in the previous study in terms of its simplicity and its ability to be implemented independently from data analysis. The weighting method used in this study has inherent characteristics that improve transparency. Transparency in the weighting method in the context of MCDM refers to the degree to which the decision-making process and the assignment of weights to criteria are clear, understandable, and explicit. The concept refers to the degree to which experts' preferences and judgments can be accurately understood and communicated to others involved in the decision-making process affected by these decisions. Furthermore, the weighting method used in this study allows for flexibility in incorporating diverse perspectives and values. In contrast, the weighting method used in the previous study may have limited flexibility in incorporating expert preferences. Unlike the previous study, this study has successfully addressed the problems of vagueness, uncertainty, and hesitancy that arise from data or expert preferences in the weighting method. Nevertheless, the present study demonstrates a broad range of options in the utilisation of FSs with the weighting method. Therefore, the present study meets all seven points (100%), while the previous study only met two points (28.6%) of the points under the weighting perspective.

In terms of the benchmarking perspective, the present study does not have the capability to ascertain the relative significance of evaluation matrices. Hence, the present study meets seven of the eight points, representing 87.5% of the total points. The previous study's implementation of the TOPSIS method did not determine the relative importance of evaluation matrices. Furthermore, the previous study failed to address the dynamic changes that occur when considering the consequences of boundary values. The previous study required the use of the TOPSIS method, which involved normalizing the data and calculating distances. In addition, the previous study also did not employ any FSs with TOPSIS, which lacked the wide range of options to effectively handle the vagueness, uncertainty, and hesitancy issues. Hence, the previous study did not fulfill any point under benchmarking perspective.

Overall, the present study successfully obtains 94.4% of the total points, specifically 17 out of 18 points, with a shortfall of only 5.6% represented by 1 point. By contrast, the previous study achieved a score of only 27.8% by meeting 5 out of 18 points, failing to meet the remaining 13 points, accounting for 72.2% of the total. The aforementioned findings suggest that the present study has successfully devised a rigorous approach to rank HMDTML alternatives and ascertain the most efficient model by considering key comparative points.

6 Practical Implication

Utilizing MCDM methods to rank HMDTML models for COVID-19 diagnosis using CXR images ensures a systematic, objective, and flexible approach to model selection. The practical implications of this approach are as follows:

First, the proposed method enhanced the process of decision making. MCDM methods provide a systematic approach for comparing and ranking models using multiple criteria, thereby ensuring the decision-making process is robust, transparent, and objective.

Second, the utilization of MCDM weighting method enables the prioritization of criteria by assigning weights according to their importance within a specific context. Healthcare providers have the ability to prioritize specific criteria, such as accuracy or recall, based on their specific requirements.

Third, the MCDM approach can be easily updated to incorporate new data or criteria, thereby maintaining the relevance and up to date of model rankings.

Fourth, the structured nature of MCDM method enables improved stakeholder communication. The clear rationale for selecting a specific model over others can inspire confidence in patients, clinicians, and other stakeholders.

Fifth, the outcomes of MCDM can serve as valuable feedback for data scientists and model developers. Understanding the position of their models in relation to various criteria can help them determine where to concentrate their efforts for improvement.

Sixth, healthcare providers can enhance patient outcomes by selecting the most effective model, which is determined by evaluating multiple criteria for accurate diagnoses.

Seventh, healthcare institutions can optimize resource allocation by identifying the most efficient and effective models, enabling them to allocate server space or computational power accordingly.

7 Study Limitations

The proposed method exhibits certain limitations. First, the conversion of the EDM and dynamic localisation decision matrix into a PSVNHS-EDM and fuzzy dynamic localisation decision matrix, respectively, is constrained to the application of a single fuzzy method, specifically, PSVNHS. Second, a single aggregation operator and scoring function were utilised to establish the ultimate weighting through PSVNH-FWZIC and the benchmarking outcomes through PSVNH-DLBD. Third, the utilisation of FS with DBLD was deemed necessary due to imprecise and unclear data. The issue of missing information in a decision matrix with respect to objective values has not been taken into account.

8 Conclusion

This study combined DLBD with FWZIC under the PSVNHF environment to benchmark the HMDTML models based on seven evaluation criteria. The methodology process started with the selection and preprocessing of the 48 HMDTML models, including 4 multideep transfer learning models and 12 multimachine learning models trained on 936 images, including 468 cases of positive COVID-19 patients and 468 cases of normal individuals, from two datasets. This was followed by the development of the HMDTML models and the construction of a uniform dynamic decision matrix. Then, PSVNH–FWZIC was formulated to estimate the weight of the evaluation criteria. Finally, the constructed matrix and the resulting weight values are fed to PSVNH–DLBD to the benchmark process.

The sensitivity of the proposed method to variations in the weights of the criteria has been demonstrated, and the mean of SCC values indicated a moderate positive association. According to our results and two types of comparisons anaylsis, DBLD can generate a more reasonable ranking of models with objective values than other existing MCDM techniques.

The findings of this study suggest that there are several possible directions for future exploration. (i) applying different fuzzy sets (e.g. probabilistic dual hesitation FS, Fermatean fuzzy soft sets or q-rung orthopair fuzzy soft sets) to the EDM and dynamic localisation decision matrix, (ii) using other aggregation and score functions with PSVNH–FWZIC and PSVNH–FDOSM, (iii) examining other numbers of scales [Eq. (14)] on the basis of the fuzzy environment standard, (vi) examining the missing information problem with the DBLD method, (v) utilising a case study with a missing information problem and examining how FDOSM can address such a problem, and (vi) extending FDOSM under the PSVNH environment.

References

Abdul, S. A. K., Suppiah, S., Ibrahim, R. M., Nasir, N. N. M., Bahari, N., Hassan, H. A., Saini, S. M., Ibrahim, I., Zakaria, M. H., & Mahmud, R. (2021). Prevalence of ultrasound-diagnosed non alcoholic fatty liver disease among rural indigenous population in malaysian and its association with biochemical and anthropometric measures. *Malaysian Journal of Medicine and Health Sciences*, *17*(3), 88–97.

Ahuja, S., Panigrahi, B. K., Dey, N., Rajinikanth, V., & Gandhi, T. K. (2021). Deep transfer learning-based automated detection of COVID-19 from lung CT scan slices. *Applied Intelligence*, *51*(1), 571–585. https://doi.org/10.1007/s10489-020-01826-w

Alamleh, A., Baqer, M. J., Jasim, A. N., & Al-Samarraay, M. S. (2022). Multi-Attribute Decision-Making for Intrusion Detection Systems: A Systematic Review. *International Journal of Information Technology and Decision Making*. https://doi.org/10.1142/S021962202230004X

Alqudah, A. M., Qazan, S., & Alqudah, A. (2020). Automated Systems for Detection of COVID-19 Using Chest X-ray Images and Lightweight Convolutional Neural Networks. *Emergency Radiology*, *4*(1), 54–67. https://doi.org/10.21203/rs.3.rs-24305/v1

Ammas, A. S. A., Mahmud, R., Hassan, H. A., Ibrahim, I., & Mohammed, S. S. (2022). An assessment of plantar fascia with ultrasound findings in patients with plantar fasciitis: a systematic review. *Journal of Ultrasound*. https://doi.org/10.1007/s40477-022-00712-0

Baqer, M. J., Izhar, M. A. M., & Abbas, I. T. (2023). A Decision Modeling Approach for Data Acquisition Systems of the Vehicle Industry Based on Interval-Valued Linear Diophantine Fuzzy Set. *International Journal of Information Technology & Decision Making*. https://doi.org/10.1142/S0219622023500487

Bouchareb, Y., Moradi Khaniabadi, P., Al Kindi, F., Al Dhuhli, H., Shiri, I., Zaidi, H., & Rahmim, A. (2021). Artificial intelligence-driven assessment of radiological images for COVID-19. *Computers in Biology and Medicine*, *136*. https://doi.org/10.1016/j.compbiomed.2021.104665

Cai, C., Gou, B., Khishe, M., Mohammadi, M., Rashidi, S., Moradpour, R., & Mirjalili, S. (2023). Improved deep convolutional neural networks using chimp optimization algorithm for Covid19 diagnosis from the X-ray images. *Expert Systems with Applications*, *213*, 119206. https://doi.org/10.1016/j.eswa.2022.119206

Caobelli, F. (2020). Artificial intelligence in medical imaging: Game over for radiologists? *European Journal of Radiology*, *126*. https://doi.org/10.1016/j.ejrad.2020.108940

Cohen, J. P., Morrison, P., Dao, L., Roth, K., Duong, T. Q., & Ghassemi, M. (2020). COVID-19 Image Data Collection: Prospective Predictions Are the Future. *Journal of Machine Learning for Biomedical Imaging*, *2020*, 2–3. https://doi.org/10.48550/arxiv.2006.11988

Dalic, I., Stevic, Ž., Karamasa, C., & Puška, A. (2020). A novel integrated fuzzy piprecia-interval rough saw model: Green supplier selection. *Decision Making: Applications in Management and Engineering*, *3*(1), 126–145. https://doi.org/10.31181/dmame2003114d

Deveci, M., Pamucar, D., & Delen, D. (2023). Performance assessment of sustainable transportation in the shipping industry using a q-rung orthopair fuzzy rough sets-based decision making methodology. *Expert Systems with Applications*, 119958. https://doi.org/10.1016/J.ESWA.2023.119958

Fusco, R., Grassi, R., Granata, V., Setola, S. V., Grassi, F., Cozzi, D., Pecori, B., Izzo, F., & Petrillo, A. (2021). Artificial intelligence and covid-19 using chest ct scan and chest x-ray images: Machine learning and deep learning approaches for diagnosis and treatment. *Journal of Personalized Medicine*, *11*(10), 993. https://doi.org/10.3390/jpm11100993

Goel, T., Murugan, R., Mirjalili, S., & Chakrabartty, D. K. (2021). OptCoNet: an optimized convolutional neural network for an automatic diagnosis of COVID-19. *Applied Intelligence*, *51*(3), 1351–1366. https://doi.org/10.1007/s10489-020-01904-z

He, X., Wang, S., Chu, X., Shi, S., Tang, J., Liu, X., Yan, C., Zhang, J., & Ding, G. (2021). Automated Model Design and Benchmarking of Deep Learning Models for COVID-19 Detection with Chest CT Scans. *35th AAAI Conference on Artificial Intelligence, AAAI 2021*, *6A*, 4821–4829. https://doi.org/10.1101/2020.06.08.20125963

Hu, T., Khishe, M., Mohammadi, M., Parvizi, G.-R., Taher Karim, S. H., & Rashid, T. A. (2021). Real-time COVID-19 diagnosis from X-Ray images using deep CNN and extreme learning machines stabilized by chimp optimization algorithm. *Biomedical Signal Processing and Control*, *68*, 102764. https://doi.org/10.1016/j.bspc.2021.102764

Jain, R., Gupta, M., Taneja, S., & Hemanth, D. J. (2021). Deep learning based detection and analysis of COVID-19 on chest X-ray images. *Applied Intelligence*, *51*(3), 1690–1700. https://doi.org/10.1007/s10489-020-01902-1

Kaplan, S. A. (2016). Re: EAU Guidelines on the Assessment of Non-Neurogenic Male Lower Urinary Tract Symptoms Including Benign Prostatic Obstruction. *The Journal of Urology*, *196*(6), 1712–1714. https://doi.org/10.1016/j.juro.2016.09.017

Kermany, D. S., Goldbaum, M., Cai, W., Valentim, C. C. S., Liang, H., Baxter, S. L., McKeown, A., Yang, G., Wu, X., Yan, F., Dong, J., Prasadha, M. K., Pei, J., Ting, M., Zhu, J., Li, C., Hewett, S., Dong, J., Ziyar, I., … Zhang, K. (2018). Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning. *Cell*, *172*(5), 1122-1131.e9. https://doi.org/10.1016/j.cell.2018.02.010

Khammas, A. S. A., Hassan, H. A., Ibrahim, R. M., Nasir, N. N. M., Bahari, N., Suppiah, S., Saini, S. M., Ibrahim, I., Zakaria, M. H., & Mahmud, R. (2020). Sonographic assessment of renal size and its correlation with anthropometric measures among indigenous population in peninsular Malaysia. *Malaysian Journal of Medicine and Health Sciences*, *16*(4), 138–145.

Khan, I. U., & Aslam, N. (2020). A deep-learning-based framework for automated diagnosis of COVID-19 using X-ray images. *Information (Switzerland)*, *11*(9). https://doi.org/10.3390/INFO11090419

Khishe, M., Caraffini, F., & Kuhn, S. (2021). Evolving Deep Learning Convolutional Neural Networks for Early COVID-19 Detection in Chest X-ray Images. *Mathematics*, *9*(9), 1002. https://doi.org/10.3390/math9091002

Liao, T. W. (2015). Two interval type 2 fuzzy TOPSIS material selection methods. *Materials and Design*, *88*, 1088–1099. https://doi.org/10.1016/j.matdes.2015.09.113

Liu, Y. H. (2018). Feature Extraction and Image Recognition with Convolutional Neural Networks. *Journal of Physics: Conference Series*, *1087*(6), 62032. https://doi.org/10.1088/1742-6596/1087/6/062032

Loey, M., Smarandache, F., & Khalifa, N. E. M. (2020). Within the lack of chest COVID-19 X-ray dataset: A novel detection model based on GAN and deep transfer learning. *Symmetry*, *12*(4), 651. https://doi.org/10.3390/SYM12040651

López-Cabrera, J. D., Orozco-Morales, R., Portal-Diaz, J. A., Lovelle-Enríquez, O., & Pérez-Díaz, M. (2021). Current limitations to identify COVID-19 using artificial intelligence with chest X-ray imaging. *Health and Technology*, *11*(2), 411–424. https://doi.org/10.1007/s12553-021-00520-2

Maghdid, H., Asaad, A. T., Ghafoor, K. Z. G., Sadiq, A. S., Mirjalili, S., & Khan, M. K. K. (2021). Diagnosing COVID-19 pneumonia from x-ray and CT images using deep learning and transfer learning algorithms. *Spiedigitallibrary.Org*, 26. https://doi.org/10.1117/12.2588672

Makris, A., Kontopoulos, I., & Tserpes, K. (2020). COVID-19 detection from chest X-ray images using deep learning and convolutional neural networks. *ACM International Conference Proceeding Series*, 60–66. https://doi.org/10.1145/3411408.3411416

Malik, Y. S., Sircar, S., Bhat, S., Ansari, M. I., Pande, T., Kumar, P., Mathapati, B., Balasubramanian, G., Kaushik, R., Natesan, S., Ezzikouri, S., El Zowalaty, M. E., & Dhama, K. (2021). How artificial intelligence may help the Covid-19 pandemic: Pitfalls and lessons for the future. *Reviews in Medical Virology*, *31*(5), 1–11. https://doi.org/10.1002/rmv.2205

Mathew, M., Chakrabortty, R. K., & Ryan, M. J. (2020). A novel approach integrating AHP and TOPSIS under spherical fuzzy sets for advanced manufacturing system selection. *Engineering Applications of Artificial Intelligence*, *96*. https://doi.org/10.1016/j.engappai.2020.103988

Medhat, M. A., & El Kassas, M. (2020). Letter regarding "High rates of 30-day mortality in patients with cirrhosis and COVID-19." *Journal of Hepatology*, *73*(6), 1569–1570. https://doi.org/10.1016/j.jhep.2020.06.023

Minaee, S., Kafieh, R., Sonka, M., Yazdani, S., & Jamalipour Soufi, G. (2020). Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning. *Medical Image Analysis*, *65*. https://doi.org/10.1016/j.media.2020.101794

Mohammed, M., Abdulkareem, K. H., Al-Waisy, A. S., Mostafa, S. A., Al-Fahdawi, S., Dinar, A. M., Alhakami, W., Baz, A., Al-Mhiqani, M. N., Alhakami, H., Arbaiy, N., Maashi, M. S., Mutlag, A. A., Garcia-Zapirain, B., & De La Torre Diez, I. (2020). Benchmarking Methodology for Selection of Optimal COVID-19 Diagnostic Model Based on Entropy and TOPSIS Methods. *IEEE Access*, *8*, 99115–99131. https://doi.org/10.1109/ACCESS.2020.2995597

Mohammed, R. T., Yaakob, R., Sharef, N. M., Abdullah, R. H., Albahri, O. S., & Abdulkareem, K. H. (2022). Determining Importance of Many-Objective Optimisation Competitive Algorithms Evaluation Criteria Based on a Novel Fuzzy-Weighted Zero-Inconsistency Method. *International Journal of Information Technology and Decision Making*, *21*(1), 195–241. https://doi.org/10.1142/S0219622021500140

Mohammed, Z. K., Aris, H. B., Roza, S. (2023). Bitcoin network-based anonymity and privacy model for metaverse implementation in Industry 5.0 using linear Diophantine fuzzy sets. *Annals of Operations Research*. https://doi.org/10.1007/s10479-023-05421-3

Nanni, L., Ghidoni, S., & Brahnam, S. (2017). Handcrafted vs. non-handcrafted features for computer vision classification. *Pattern Recognition*, *71*, 158–172. https://doi.org/10.1016/j.patcog.2017.05.025

Nayak, S. R., Nayak, D. R., Sinha, U., Arora, V., & Pachori, R. B. (2021). Application of deep learning techniques for detection of COVID-19 cases using chest X-ray images: A comprehensive study. *Biomedical Signal Processing and Control*, *64*. https://doi.org/10.1016/j.bspc.2020.102365

Nkurunziza, J. M. V., Udahemuka, J. C., Dusenge, J. B., & Umutesi, F. (2022). Overview of Trending Medical Technologies. *Global Clinical Engineering Journal*, *4*(3), 16–46. https://doi.org/10.31354/globalce.v4i3.142

Nour, M., Cömert, Z., & Polat, K. (2020). A Novel Medical Diagnosis model for COVID-19 infection detection based on Deep Features and Bayesian Optimization. *Applied Soft Computing*, *97*. https://doi.org/10.1016/j.asoc.2020.106580

Ozturk, T., Talo, M., Yildirim, E. A., Baloglu, U. B., Yildirim, O., & Rajendra Acharya, U. (2020). Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Computers in Biology and Medicine*, *121*. https://doi.org/10.1016/j.compbiomed.2020.103792

Pamucar, D., Yazdani, M., Obradovic, R., Kumar, A., & Torres-Jiménez, M. (2020). A novel fuzzy hybrid neutrosophic decision-making approach for the resilient supplier selection problem. *International Journal of Intelligent Systems*, *35*(12), 1934–1986. https://doi.org/10.1002/int.22279

Panigrahi, S., Nanda, A., & Swarnkar, T. (2021). A Survey on Transfer Learning. *Smart Innovation, Systems and Technologies*, *194*, 781–789. https://doi.org/10.1007/978-981-15-5971-6_83

Panwar, H., Gupta, P. K., Siddiqui, M. K., Morales-Menendez, R., & Singh, V. (2020). Application of deep learning for fast detection of COVID-19 in X-Rays using nCOVnet. *Chaos, Solitons and Fractals*, *138*. https://doi.org/10.1016/j.chaos.2020.109944

Pathan, S., Siddalingaswamy, P. C., & Ali, T. (2021). Automated Detection of Covid-19 from Chest X-ray scans using an optimized CNN architecture. *Applied Soft Computing*, *104*. https://doi.org/10.1016/j.asoc.2021.107238

Pereira, R. M., Bertolini, D., Teixeira, L. O., Silla, C. N., & Costa, Y. M. G. (2020). COVID-19 identification in chest X-ray images on flat and hierarchical classification scenarios. *Computer Methods and Programs in Biomedicine*, *194*. https://doi.org/10.1016/j.cmpb.2020.105532

Qumar, A., Sah, B., Singh, A. R., Deng, Y., He, X., Kumar, P., & Bansal, R. C. (2022). Evaluation of agriculture-food 4.0 supply chain approaches using Fermatean probabilistic hesitant-fuzzy sets based decision making model. *Applied Soft Computing*, 110170. https://doi.org/10.1016/J.ASOC.2023.110170

Qumar, S., Kou, G., Lu, Y., Peng, Y., & Shi, Y. (2023). A comparative study of evaluating and benchmarking sign language recognition system-based wearable sensory devices using a single fuzzy set. *Knowledge-Based Systems*, 110519. https://doi.org/10.1016/J.KNOSYS.2023.110519

Qassim, F., Tzeng, G. H., Chiang, C. H., & Li, C. W. (2022). Integrated sustainable transportation modelling approaches for electronic passenger vehicle in the context of industry 5.0. *Journal of Innovation and Knowledge*, *7*(4), 100277. https://doi.org/10.1016/j.jik.2022.100277

Ramleh, S., Yatim, K., Zulzalil, H., Osman, M. H. (2022). Review of healthcare industry 4.0 application-based blockchain in terms of security and privacy development attributes: Comprehensive taxonomy, open issues and challenges and recommended solution. *Journal of Network and Computer Applications*, *209*, 103529. https://doi.org/10.1016/J.JNCA.2022.103529

Rahaman, M. M., Li, C., Yao, Y., Kulwa, F., Rahman, M. A., Wang, Q., Qi, S., Kong, F., Zhu, X., & Zhao, X. (2020). Identification of COVID-19 samples from chest X-Ray images using deep learning: A comparison of transfer learning approaches. *Journal of X-Ray Science and Technology*, *28*(5), 821–839. https://doi.org/10.3233/XST-200715

Ravi, V., Narasimhan, H., Chakraborty, C., & Pham, T. D. (2021). Deep learning-based meta-classifier approach for COVID-19 classification using CT scan and chest X-ray images. *Multimedia Systems*. https://doi.org/10.1007/s00530-021-00826-1

Rong, X., Yang, L., Chu, H., & Fan, M. (2020). Effect of delay in diagnosis on transmission of COVID-19. *Mathematical Biosciences and Engineering*, *17*(3), 2725–2740. https://doi.org/10.3934/mbe.2020149

Rozi, H. A., Kou, G., Malik, R. Q., & Jasim, A. N. (2022). Integration of FDOSM and FWZIC Under Homogeneous Fermatean Fuzzy Environment: A Prioritization of COVID-19 Patients for Mesenchymal Stem Cell Transfusion. International Journal of Information Technology and Decision Making. https://doi.org/10.1142/S0219622022500511

Rubin, G. D., Ryerson, C. J., Haramati, L. B., Sverzellati, N., Kanne, J. P., Raoof, S., Schluger, N. W., Volpi, A., Yim, J. J., Martin, I. B. K., Anderson, D. J., Kong, C., Altes, T., Bush, A., Desai, S. R., Goldin, O., Mo Goo, J., Humbert, M., Inoue, Y., … Leung, A. N. (2020). The role of chest imaging in patient management during the covid-19 pandemic: A multinational consensus statement from the fleischner society. *Radiology*, *296*(1), 172–180. https://doi.org/10.1148/radiol.2020201365

Saffari, A., Khishe, M., Mohammadi, M., Hussein Mohammed, A., & Rashidi, S. (2022). DCNN-FuzzyWOA: Artificial Intelligence Solution for Automatic Detection of COVID-19 Using X-Ray Images. *Computational Intelligence and Neuroscience*, *2022*, 1–11. https://doi.org/10.1155/2022/5677961

Şahin, R., & Altun, F. (2020). Decision making with MABAC method under probabilistic single-valued neutrosophic hesitant fuzzy environment. *Journal of Ambient Intelligence and Humanized Computing*, *11*(10), 4195–4212. https://doi.org/10.1007/s12652-020-01699-4

Salih, M. M., Zadan, B., & Zadan, K. (2020). Fuzzy decision by opinion score method. *Applied Soft Computing Journal*, *96*. https://doi.org/10.1016/j.asoc.2020.106595

Sethy, P. K., Behera, S. K., Ratha, P. K., & Biswas, P. (2020). Detection of coronavirus disease (COVID-19) based on deep features and support vector machine. *International Journal of Mathematical, Engineering and Management Sciences*, *5*(4), 643–651. https://doi.org/10.33889/IJMEMS.2020.5.4.052

Shao, S., Zhang, X., Li, Y., & Bo, C. (2018). Probabilistic single-valued (Interval) neutrosophic hesitant fuzzy set and its application in multi-attribute decision making. *Symmetry*, *10*(9). https://doi.org/10.3390/sym10090419

Shibly, K. H., Dey, S. K., Islam, M. T. U., & Rahman, M. M. (2020). COVID faster R–CNN: A novel framework to Diagnose Novel Coronavirus Disease (COVID-19) in X-Ray images. *Informatics in Medicine Unlocked*, *20*. https://doi.org/10.1016/j.imu.2020.100405

Toğaçar, M., Ergen, B., & Cömert, Z. (2020). COVID-19 detection using deep learning models to exploit Social Mimic Optimization and structured chest X-ray images using fuzzy color and stacking approaches. *Computers in Biology and Medicine*, *121*. https://doi.org/10.1016/j.compbiomed.2020.103805

Wang, X., Gong, C., Khishe, M., Mohammadi, M., & Rashid, T. A. (2022). Pulmonary Diffuse Airspace Opacities Diagnosis from Chest X-Ray Images Using Deep Convolutional Neural Networks Fine-Tuned by Whale Optimizer. *Wireless Personal Communications*, *124*(2), 1355–1374. https://doi.org/10.1007/s11277-021-09410-2

Yadav, N., Ahuja, A., Ashish, M., Nijhawan, R., & Sharma, B. (2022). A Modern Replica for COVID-19 Pestilential Disease Identification. *Proceedings of 3rd International Conference on Intelligent Engineering and Management, ICIEM 2022*, 81–88. https://doi.org/10.1109/ICIEM54221.2022.9853060

Zaidan, A. A., Alsattar, H. A., Sammah, S., & Hajiaghaei-Keshteli, M. (2023). Uncertainty Decision Modeling Approach for Control Engineering Tools to Support Industrial Cyber-Physical Metaverse Smart Manufacturing Systems. *IEEE Systems Journal*, 1–12. https://doi.org/10.1109/JSYST.2023.3266842

Zhou, L. P., Dong, J. Y., & Wan, S. P. (2019). Two new approaches for multi-attribute group decision-making with interval-valued neutrosophic frank aggregation operators and incomplete weights. *IEEE Access*, *7*, 102727–102750. https://doi.org/10.1109/ACCESS.2019.2927133