Journal of Medical Systems (2019) 43: 212
https://doi.org/10.1007/s10916-019-1338-x

SYSTEMS-LEVEL QUALITY IMPROVEMENT

Multiclass Benchmarking Framework for Automated Acute Leukaemia Detection and Classification Based on BWM and Group-VIKOR

M. A. Alsalem[1,2] · A. A. Zaidan[1] · B. B. Zaidan[1] · O. S. Albahri[1] · A. H. Alamoodi[1] · A. S. Albahri[3] · A. H. Mohsin[4] · K. I. Mohammed[1]

Received: 4 March 2019 / Accepted: 13 May 2019 / Published online: 1 June 2019

Abstract

This paper aims to assist the administration departments of medical organisations in making the right decision on selecting a suitable multiclass classification model for acute leukaemia. In this paper, we proposed a framework that will aid these departments in evaluating, benchmarking and ranking available multiclass classification models for the selection of the best one. Medical organisations have continuously faced evaluation and benchmarking challenges in such endeavour, especially when no single model is superior. Moreover, the improper selection of multiclass classification for acute leukaemia model may be costly for medical organisations. For example, when a patient dies, one such organisation will be legally or financially sued for incidents in which the model fails to fulfil its desired outcome. With regard to evaluation and benchmarking, multiclass classification models are challenging processes due to multiple evaluation and conflicting criteria. This study structured a decision matrix (DM) based on the crossover of 2 groups of multi-evaluation criteria and 22 multiclass classification models. The matrix was then evaluated with datasets comprising 72 samples of acute leukaemia, which include 5327 gens. Subsequently, multi-criteria decision-making (MCDM) techniques are used in the benchmarking and ranking of multiclass classification models. The MCDM used techniques that include the integrated BWM and VIKOR. BWM has been applied for the weight calculations of evaluation criteria, whereas VIKOR has been used to benchmark and rank classification models. VIKOR has also been employed in two decision-making contexts: individual and group decision making and internal and external group aggregation. Results showed the following: (1) the integration of BWM and VIKOR is effective at solving the benchmarking/selection problems of multiclass classification models. (2) The ranks of classification models obtained from internal and external VIKOR group decision making were almost the same, and the best multiclass classification model based on the two was 'Bayes. Naive Byes Updateable' and the worst one was 'Trees.LMT'. (3) Among the scores of groups in the objective

This article is part of the Topical Collection on *Systems-Level Quality Improvement*

✉ A. A. Zaidan
aws.alaa@gmail.com; aws.alaa@fskik.upsi.edu.my

M. A. Alsalem
mohammed.asum@gmail.com

B. B. Zaidan
bilalbahaa@fskik.upsi.edu.my

O. S. Albahri
osamahsh89@gmail.com

A. H. Alamoodi
Abdullahalamood@outlook.com

A. S. Albahri
ahmed.bahri1978@gmail.com

A. H. Mohsin
ali_hadi182@yahoo.com

K. I. Mohammed
khalid_ib81@yahoo.com

[1] Department of Computing, Universiti Pendidikan Sultan Idris, Tanjong Malim, Perak, Malaysia

[2] Department of Management Information System, College of Administration and Economic, University of Mosul, Mosul, Iraq

[3] College of Engineering, University of Information Technology and Communications, Baghdad, Iraq

[4] Republic of Iraq-Presidency of Ministries - Establishment of Martyrs, Baghdad, Iraq

validation, significant differences were identified, which indicated that the ranking results of internal and external VIKOR group decision making were valid.

Keywords Classification · Acute leukaemia · BWM · VIKOR · Multiclass evaluation · Benchmarking

Introduction

Medical informatics is the intersection of information science, computer science, and health care [1–10]. This field deals with the resources, devices, and methods required to optimize the acquisition, storage, retrieval, and use of information in health [11–24]. The decisions of the administration departments of medical organisations are critical, particularly decisions regarding the selection of automated solutions for the diagnosis and detection of complex diseases, such as acute leukaemia [25]. The importance of selecting appropriate automated solutions can be attributed to their extensive use [26]. Automated solutions based on artificial intelligence techniques can provide rapid acute leukaemia diagnosis and classification and increase the reliability and accuracy of diagnostic results [26–32]. Many physicians, cancer treatment centres and hospitals have started using automated models for acute leukaemia classification to address the several potential limitations of manual analysis [26, 29, 30]. However, despite the increasing number of automated classification models, finding models that deliver highly accurate results in a short time and without error remains challenging [33]. Therefore, the administration departments of health organisations have been facing difficulties in evaluating and benchmarking automated classification models for acute leukaemia and determining the best model, especially when no single model is superior [29, 33, 34]. Moreover, evaluating and comparing different classification models is difficult in the presence of multiple evaluation criteria [35, 36]. Given the existence of different classification models for acute leukaemia, the health sector has difficulty deciding which model should be used. The required processes for tasks related to the evaluation and benchmarking of automated classification models for dangerous medical cases are crucial to the identification of the classification model that delivers the best results [27]. These processes are crucial because the selection of an incorrect classification model can lead to the loss of a patient's life, legal accountability and even financial costs for the health organisations. For example, when a model incorrectly identifies non-cancer cells as cancerous in a patient, the surgery and diagnostic tests the patient have to undergo may pose adverse effects on his or her mental health. Conversely, when a model incorrectly identifies cancer cells as non-cancerous, the disease remains untreated, and the patient may die as a result. Both cases have a negative impact on the reputation and performance of healthcare organisations. Therefore, determining the most efficient technique for selecting a suitable classification model for acute leukaemia is necessary. Given that these models are not cheap, as well as related to the medical aspect for humans, they must be evaluated and benchmarked [35]. The procedures related to the multiclass classification of acute leukaemia through evaluation and benchmarking remains challenging [29]. The tasks involved in the evaluation and benchmarking automated models for acute leukaemia are difficult decision-making tasks and requires numerous measurements [34]. Two basic sets of criteria are commonly utilised in the evaluation and benchmarking of acute leukaemia multiclass classification modes: (1) time complexity and (2) reliability group. The first group for reliability has a set of sub-criteria (TP, TN, FP, FN, ave-accuracy, precisionμ, precision_M, recall_M, fscore and error rate) [37, 38]. Snousy et al. considered the main requirements for the best classification model in terms of accuracy [33], and nine classification models based on accuracy criterion were compared in their study. Despite the importance of the remaining criteria [39–41], several studies [32, 42–46] adopted the classification accuracy criterion for the evaluation and benchmarking of classification models. However, the quality assessment of acute leukaemia classification models requires additional attention. In the same context, some other aspects must be considered in the evaluation processes [33]. According to Rawat et al., although accuracy is the most widely used metric, each class of aspect is considered with equal importance, and the differences among the types of classes are neglected [32]. However, in real cases, particularly those related to medicine, the distinction among certain classified classes is important. In [47–49], True Positive, True Negative, False Positive and False Negative sensitivity were used as key criteria for evaluation and benchmarking, but other requirements that might have an impact on classification performance were neglected. In [35], the calculation of time complexity was found to be time consuming for classification. High computational cost causes the slowdown of classification [50]. Misha et al. indicated that the dataset size should be considered in the classification task because a large dataset affects processing time; this condition is known as time complexity [35]. Ludwig et al. stated that in the scope of cancer data analysis, speed and accuracy are the main aspects that must be considered in the evaluation of the efficiency of classification models [51]. Classification tasks are considered good if the results with low computational time are delivered and classification accuracy is simultaneously improved [52]. In other words, the main requirements that must be considered

when developing any acute leukaemia multiclass classification model are as follows: (1) time complexity and (2) reliability. Reliability should have a high rate, and the time complexity for conducting the output should be low [52]. However, these requirements are competing requirements [53]; that is, high reliability cannot simultaneously be obtained with low time complexity. Thus, the developers usually focus on either increasing reliability or decreasing time complexity. If a highly reliable multiclass classification model is required, then time must be sacrificed, and vice versa. The trade-off and conflict among the evaluation criteria are reflected on the evaluation and benchmarking process. This situation leads to conflicts among criteria in the comparison, and the benchmarking process is affected. Consequently, benchmarking among multiple criteria is difficult with trade-off and conflict [54]. Reliability and time complexity should be measured in the evaluation of any classification model. However, current approaches for comparing novel and previous models in all the reviewed studies do not focus on the evaluation and benchmarking criteria; they only emphasised the evaluation aspect and neglected the rest because they are not sufficiently flexible to deal with the conflict or trade-off among the various criteria [33]. Conflict and trade-off are considered the first issue faced by the evaluation and benchmarking of multiclass classification models. The second issue is the importance of each criterion. Acute leukaemia evaluation in terms of multiclass classification models involves a set of criteria, and the importance of each criterion is distinct and depends on the objectives of the developed model. That is, the importance of one of the evaluation criteria might be boosted in exchange for the low importance of another criterion based on model objectives [34]. Therefore, trade-off and conflict exist between evaluation and benchmarking criteria due to importance differences of each criterion in different models [55]. The third issue emerges when the benchmarking process is conducted on the basis of simultaneous multiple criteria and sub-criteria [56–58]. This approach is considered to be difficult due to the trade-off among the criteria and their various importance; however, the reliability of a criteria set indicates that the values depend on the confusion matrix containing four parameters: True Positive, False Positive, True Negative and False Negative [47, 59]. The four parameters are prone to lose values in experiments, affecting the remaining values of other criteria in the reliability group. Despite the criticism with respect to these parameters, the studies still used these parameters for the evaluation of multiclass classification models [56–58, 60]. By contrast, the current evaluation and benchmarking tools have limitations. These tools cannot entirely cover the required measurements by the multiclass classification model. Moreover, these tools have limitations in terms of the overall parameter calculation of the reliability group, comparison between the two additional classification methods and matching between the classification methods because the tools cannot rank the models according to performance [61–63]. In the preceding discussion, the problem of evaluation and benchmarking process in multiclass classification models of acute leukaemia is defined as a multi-criteria problem. Therefore, an integrated and comprehensive platform covering all the aspects of performance in the evaluation and benchmarking of multiclass classification models for acute leukaemia should be developed. This integrated platform will serve as a tool that supports the decisions of the administrators of medical organisations in the evaluation and benchmarking of available alternatives and the identification of the best model. The main objective of the current paper is to propose a framework for evaluating and benchmarking multiclass classification models for acute leukaemia. The remaining parts of this article are divided into the following seven sections: the 'Related Studies' section presents related literature review. The 'Multi-criteria decision-making' section shows the theoretical background of the recommended solution. The 'Methodology' section reports the evaluation and benchmarking framework for multiclass classification models. The results and discussion are reported in the 'Results and discussion' section. 'Validation' deliberates the validation results for the proposed framework. The 'Limitations and future study' section highlights the limitations of the proposed framework and future studies. The 'Conclusion' section presents the conclusion of the research.

Related studies

The selection of a suitable classification model for acute leukaemia is considered a challenge faced by medical institutions, especially those with specialisation in cancer treatment. The essence of the challenge lies in the capacity of the selected model to allow a precise and immediate acute leukaemia classification.

Previous literature distinctly explained that classification tasks of acute leukaemia differ with respect to result accuracy provided and overall performance. Similarly [29, 33, 34], no previous classification model has been considered superior. Many studies have discussed the development of automated models for acute leukaemia analysis, as well as the way the models is used and the benefits that health organisations could gain from using them [29, 32, 34, 47, 49, 64–69]. However, studies that aimed to provide an evaluation and benchmarking of available classification models and determine the best one are limited. Existing academic literature featuring topics related to the evaluation and benchmarking of acute leukaemia multiclass classification models are scarce and scattered; some studies are only limited to the evaluation and benchmarking of one aspect of performance. In [70], automated microscopy was analysed with DM96TM. Snousy et al. compared nine

classification models under decision tree family in terms of accuracy and explored their performance in determining blood cells, then compared their accuracy with that of the manual method and XE-2100TM. The study attempted to examine experimental effects to different methods of feature selection with respect to accuracy [33]. An ALL-IDB, which is a public image dataset of peripheral blood samples for normal people and patients with leukaemia, was proposed in [27]; supervised classification and segmentation of the data were provided by the image dataset, which is particularly designed for comparing and evaluating algorithms for segmentation and classification. In [71], three automatic detection approaches for leukaemic cells were compared. The first approach is based on support vector machine, the second is based on a neural network and the third is Gaussian mixture model estimation. The comparison relied on three criteria, namely, accuracy, precision and recall. In addition to the effect of various segmentations on classification results, in [39], two classification schemes were compared in terms of segmentation quality. The first scheme is based on support vector machine, whereas the second is based on random forest. Evaluation and benchmarking methods must be utilised to cover all main requirements and substantively determine the performance and quality of classification models for acute leukaemia. In addition to reduced processing time and small error rate,*Saritha* et al. assured that the automated classification model has high accuracy and efficiency. Suitable treatment to patients can be provided with the early identification of leukaemia [52]. Despite the substantial effort in the evaluation and benchmarking of acute leukaemia classification tasks, no study has provided an integrated solution that covers the key evaluation criteria for evaluating and benchmarking multiclass classification models and helping the administrators of medical organisations and various users to determine a suitable model. This study attempts to fill the evaluation and benchmarking research gap with respect to acute leukaemia classification tasks.

Multi-criteria decision making (MCDM)

Numerous MCDM definitions are available in academic literature. However, MCDM was defined by Keeney and Raiffa [72] as decision theory extension, which is aimed to cover the decision of any multiple objectives. MCDM is used as a methodology to aid in cases, such as those assessing alternatives on individuals, which are often followed by conflicting criteria and combined into one overall appraisal [73–77]. Among the other definitions of MCDM, [78] defined MCDM as an umbrella term, which describes the collection of formal approaches. These processes decide to take explicit account of multiple criteria to assist individuals or groups exploring important decisions which matter [79–84]. Among the most well-known decision techniques, MCDM is known for its decision-making capabilities, enabling it to address complicated decision problems whilst handling multiple criteria [85, 86]. Furthermore, MCDM demonstrates a systematic method to address decision problems on the basis of multiple criteria [86–90]. The goal is to help decision makers deal with this kind of problems [91]. MCDM procedure often relies on approaches with quantitative and qualitative nature and frequently concentrates on simultaneously dealing with multiple and conflicting criteria [92, 93]. MCDM also has the capabilities to increase decision quality based on the approach via effective and rational ways more than traditional processes [94]. Furthermore, MCDM intends to acquire the following: categorise suitable alternatives among a group of available ones and rank the alternatives according to performance in decreasing order [95–99]. The last is the selection of these alternatives [100–106]. Suitable alternatives will be scored based on the previous goals. Essential terms are required in any MCDM solution, namely, the decision or evaluation matrix, which are also called decision criteria [107]. Decision matrix must be created using elements, including n criteria and m alternatives. Each criteria intersection and alternative is specified as x_ij. Therefore, matrix (x_ij) _ (m*n) is expressed as follows:

$$D = \begin{array}{c} \\ A_1 \\ A_2 \\ \vdots \\ A_m \end{array} \begin{array}{c} \begin{array}{cccc} C_1 & C_2 & \cdots & C_n \end{array} \\ \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \end{array},$$

where A_1, A_(2)…...,A_m are possible alternatives to be ranked by the decision makers (i.e. classification models); C_1,C_(2)....…,C_n are the criteria against which the performance of each alternative is evaluated and x_ij is the rating of alternative A_i with respect to criterion C_j, and W_j is the weight of criterion C_j. Special processes must be accomplished to score the alternatives. Normalisation is included in some of these processes. Maximisation indicator, addition of weights and other processes are based on the method. For example, suppose that D is the decision matrix utilised in scoring the Ai performance of the alternative, where based on C_j. Enhancing the decision-making process is important and possible by involving decision makers and stakeholders. Using appropriate decision-making methods towards handling multi-criteria problems is also necessary. Healthcare is one of the extensively utilised domains of MCDM [93, 108]. Improving decision making in healthcare is possible through a systematic method and by determining the best decision through different MCDM methods [109, 110]. Especially, many decisions in the healthcare and medical fields are complex and unstructured [108]. Numerous MCDM techniques have been developed, and the most commonly used MCDM

techniques are th best-worst method (BWM), weighted product method (WPM), hierarchical adaptive weighting (HAW), simple additive weighting (SAW), multiplicative exponential weighting (MEW), weighted sum model (WSM), analytic network process (ANP), analytic hierarchy process (AHP), technique for order of preference by similarity to ideal solution (TOPSIS) and VlseKriterijumska Optimizacija I Kompromisno Resenje' (VIKOR), which uses different notations [1, 73, 79–81, 108, 110–121]. Available MCDM techniques are diverse, and this diversity makes the selection of suitable techniques difficult. Each technique has its own limitations and strengths [81, 109, 112, 122, 123]. Thus, selecting the most suitable MCDM method is important. To the best of the our knowledge, none of the analysed methods have been used to rank multiclass classification models for acute leukaemia. In our previous work [87], we found that BWM and VIKOR are the two of the best MCDM methods.

The current study utilised 'best–worst' methods because it can provide more consistent results than AHP and other MCDM weighting methods. Moreover, the BWM-based pairwise comparisons are fewer than those in other methods [112, 124–126]. The pairwise comparison based on BWM also focuses on reference comparisons. This condition means that this comparison executes the most important preference of criterion over all the other criteria in addition to the preference of all the other criteria of least important criterion [111, 112, 127]. Conversely, MCDM methods are frequently used to rank alternatives, and the most common is VIKOR. The method utilises the approach for compromise priority for multiple response optimisation [110, 128, 129]. VIKOR is based on an aggregating function that represents 'closeness to the ideal'. The index for VIKOR ranking is based on a particular measure of 'closeness' to the ideal solution. Furthermore, VIKOR has the capability towards the ranking of the alternatives to accurately and rapidly determine the best [128]. The style for recent VIKOR studies changed, and VIKOR is usually integrated with another MCDM method. Reviewed studies identified and provided different examples for applying VIKOR with BWM to improve consistency for subjective weights. A similar integration between VIKOR with BWM realises a robust method. Given the advantages of the two methods in overcoming uncertainties associated with the problem described in [130–136], using VIKOR and BWM is easy and clear even for those with no background on MCDM [136]. Utilising VIKOR with different cases (e.g., individual and groups) has been recommended. Two main cases of decision making are basically emphasised: the first case is decision making based on a single decision maker; the second involves many decision makers and is called group decision making (GDM), in which individuals collectively select alternatives from the ones presented to them. The decision is not attributed to any single group member because of the individual and social processes, such as social influence, which contribute to the outcome. The GDM techniques systematically collect elements and combine components from experts, including their knowledge and judgement from different fields. In relation to a group case, the judgement criteria of each expert, which require subjective judgement, are provided. The same expert assigns weight for every criterion [110, 137]. Finally, evaluation and benchmarking for acute leukaemia multiclass classification suggests a need to integrate BWM and VIKOR methods. The suggestion is based on assigning weights for criteria (reliability, time complexity rate) according to BWM and on the basis of the evaluation of an expert. The utilisation of VIKOR is recommended in the ranking of multiclass classification models.

Methodology

This section introduces the evaluation and benchmarking methodology of the automated multiclass classification models. In addition, the section will introduce the procedures and steps of the proposed framework. The output ranked multiclass classification models based on the set of criteria using the BWM and VIKOR for weighting and ranking, respectively. All the overall conceptual elements of the present study are illustrated in Fig. 1.

Construction of decision matrix

Decision matrix considers the main component in the evaluation and benchmarking framework. The main parts of decision matrix are decision criteria and alternatives. In the present case, the criteria represent the metrics used for measuring the quality of multiclass classification models. The next subsection describes the procedures followed to develop and evaluate the multiclass classification models and construct the decision matrix.

Data source

The dataset proposed by [138] for acute leukaemia microarray was adopted in this study. The dataset is recognised for its popularity and usage in the academic literature and the most frequently utilised in the papers (References [139–141], which is available for the public). The dataset has three categories for acute leukaemia: acute myelogenous leukaemia (AML), ALL B cell and ALL T cell. The dataset comprises 5327 genes and 72 samples, of which 38 are AML, 9 are ALL-B and 25 are ALL-T types.

Development of multiclass classification models

Developing multiclass classification models requires a three-step process. Firstly, the target dataset, which include the

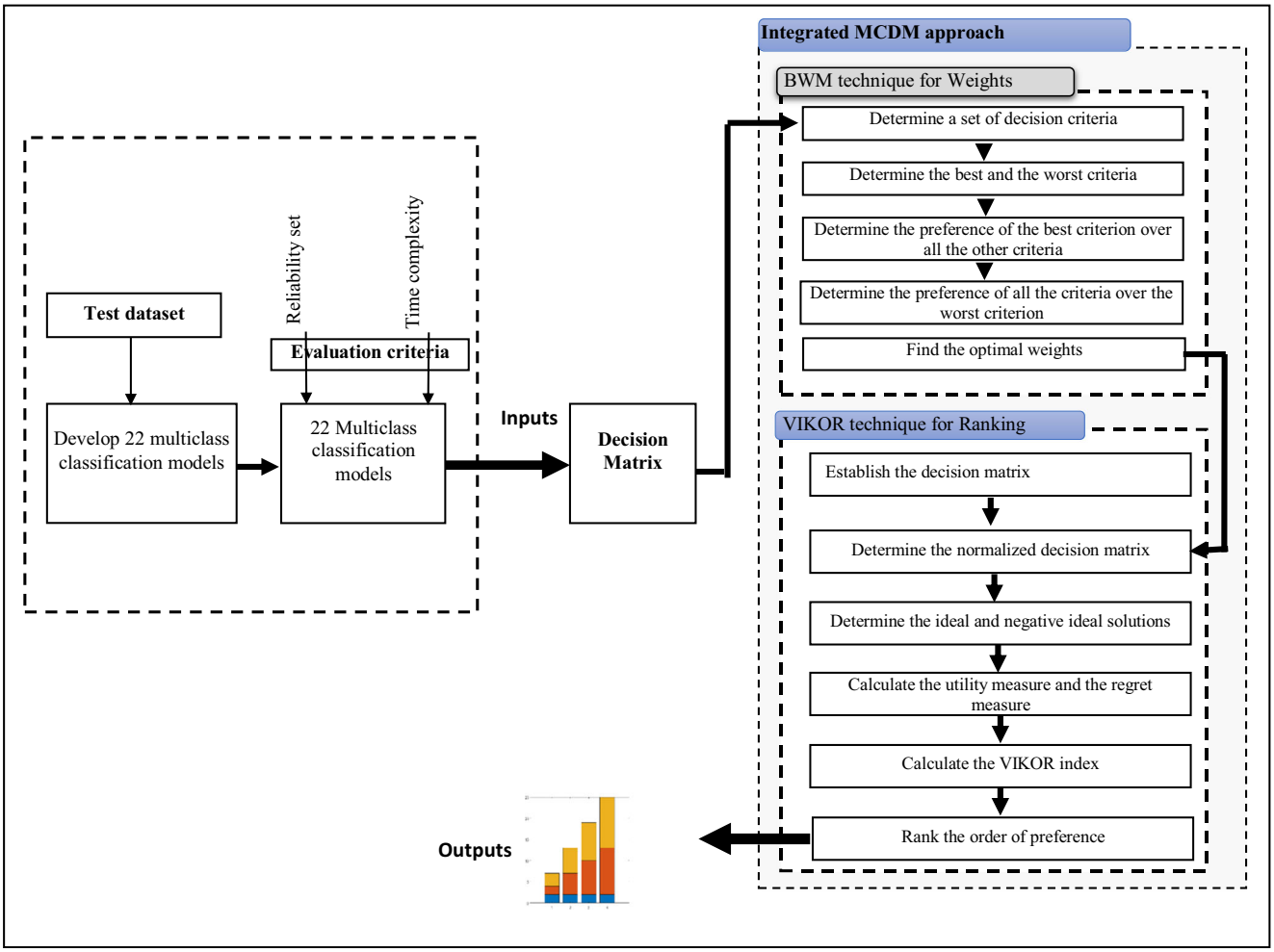

Fig. 1 Benchmarking methodology of the multiclass classification models for acute leukaemia

selection of relevant features, is prepared. Secondly, training (learning process), which involves the establishment of a class through machine learning, is achieved by analysing the instances of a training dataset. Each individual instance, should belong to a predefined class, and each instance is assumed to belong to a predefined class. Thirdly, machine learning algorithms are executed with other independent datasets, which are also known as testing datasets. This step is in line with the aim of performing machine learning estimation. If the performance for multiclass classification model appears to be 'acceptable', then the model can be utilised for future classification cases when the class label is unknown. Ultimately, the multiclass classification models, which supply an acceptable result, can be considered an acceptable multiclass classification model.

The microarray data generally contain dozens of sample sizes (small) and high dimensionality (thousands of genes). Nevertheless, the results of the classification can be affected by the genes, more specifically, few parts of these genes. This condition means that most genes have no classification value. Genes with no relevancy, apart from their negative effects on the classification performance, can cause conflict in the classification model. Moreover, given that irrelevant genes can lead to over-fitting, a positive effect can be attained by reducing the number of genes. This approach can minimise the computed input. This positive effect affects the overall performance and results for the classification [28, 142, 143]. In this study, the genes that are highly relevant with classification classes, which are known as informative genes, are selected. The chi-square (X2) [33] method was used for the individual evaluation of the features. The X2 value is computed as follows [28, 142, 143]:

$$x^2(a) = \sum_{v=V} \sum_{i=1}^{n} \frac{[A_i(a=v) - E(a=v)]2}{E_i(a=v)}, \quad (1)$$

$$x^2(a) = \sum_{v=V} \sum_{i=1}^{n} \frac{[A_i(a=v) - E(a=v)]2}{E_i(a=v)} \quad (2)$$

where V is the set of possible values for a, n is the number of class, Ai (a = V) is the number of samples in the ith class with a = v and Ei (a = v) is the expected value of Ai (a = v); Ei(a = v) = P (a = v) P (ci) N, where P(a = v) is the probability of a = v, P(ci) is the probability of one sample labelled with the ith class and N is the total number of samples [33].

A total of 22 models for multiclass classification are built based on 22 well-known machine learning algorithms available in Weka software, which have been extensively used in prior studies [33, 42, 51, 60, 142, 144–146] and demonstrated satisfactory results when used in the classification of microarray dataset. These algorithms include the following: Rule.zero, Bayes_Net, Bayes.NaiveByesUpdateable, Lazy.IBK, Meta.AdaboostM1, Meta.Bagging, Meta.filteredclassifier, Meta.logitboost, Tree.j48, REPTree, RandomTree, RandomForest, Rule. Decision Table, Rules.part, Meta.RandomCommittee, Trees.LMT, Treed.HoeffdingTree, Kstar, Functions.Smo, Functions.SIMPLE.Logistic, Byes.NaiveBayes and Decision Stump. The dataset is divided into two parts to develop multiclass classification models. The first part is utilised for training purposes, and the other is used for testing purposes. The set for training is used in training the machine learning algorithms, and the other part of the dataset (testing set) is utilised to test the trained machine learning algorithms. The test dataset is classified into three categories, namely, AML, ALL-B and ALL-T, using the 22 multiclass classifications.

Establishment and evaluation of the decision matrix

The establishment of the decision matrix is dependent on the crossover between the evaluation criteria, namely, Ave accuracy, error rate, precision_M, precision_μ, recall_M, FP, FN, TP, TN, fscore and time complexity, and the 22 developed multiclass classification models. Figure 2 presents the structure of the proposed decision matrix.

Figure 2 shows the structure of the proposed decision matrix; the top row represents the main evaluation criteria, and the first column on the left represents different developed multiclass classification models as alternatives. The values (data) in this DM denote the evaluation results of all developed multiclass classification models according to all evaluation criteria. Each multiclass classification model is evaluated based on all evaluation criteria, where the matrix of parameter, relationship of parameters, parameter behaviour and error rate represent the four sub-criteria sets in the group of reliability. Firstly, the matrix of parameter is generated (TP, TN, FN and FP), and the basic sub-criteria are represented by these parameters in the reliability group of criteria. Given that this study addressed the multiclass classification problem, one-verse all approach is used in the calculation of the reliability set of the criteria. According to these criteria, the multiclass confusion matrix is converted to three confusion matrices, and each of matrix describes the parameters for a certain class of acute leukaemia (AML, ALL-B and ALL). Based on the three confusion matrices, the remaining sub-criteria within the reliability group are calculated for each matrix by using a specific formula. Therefore, values for each multiclass classification model will be separately calculated to generate the values considering the input of the decision matrix. Finally, the calculation procedure for time complexity is based on the consumed time by two elements: the input of the dataset sample and result output. The calculation process for the sample process relies on the number and size of samples as indicated in the following equation:

$$T_{process} = T_o - T_i \tag{3}$$

where T_o is the processing time to obtain outputs, and T_i is the time of inputting the sample. The time complexity is calculated by Weka software through the experimental process. As mentioned in Section 2, the three specific issues encountered by the proposed decision matrix are as follows: (1) trade-off and conflict among the evaluation criteria, (2) multiple evaluation criteria and (3) the importance of criteria. A weight difference is observed between the main criteria and sub-criteria. MCDM is used to address this issue, as presented in the next section.

Development of the evaluation and benchmarking framework

The proposed evaluation and benchmarking framework are developed based on MCDM techniques. The framework is developed based on the integration of BWM and VIKOR for weighting and ranking the best alternatives in the proposed decision matrix and selecting the best one. The subsequent steps are presented below.

Development of evaluation and benchmarking/selection integrated methods of BWM and VIKOR using MCDM

The suitable methods for benchmarking and ranking multiclass classification models are BWM and VIKOR. The VIKOR method is a mathematical model recommended for ranking and solving specific issues related to (1) trade-off and conflict and (2) multi-evaluation criteria encountered by the proposed decision matrix. BWM is also used for weighting the criteria to solve (3) the importance of criteria in relation to the proposed decision matrix.

Accordingly, the combination of BWM and VIKOR methods is justified for benchmarking and ranking the multiclass classification models.

Calculation of the weights of criteria based on BWM method

Assigning proper weights for multi-service criteria using BWM requires several steps. The procedure for BWM includes the following steps [112, 147]:

Step 1. Determining a set of decision criteria

For BWM, the first step is to determine the criteria set, C1, C2,… Cn, which should be considered by the decision maker when selecting the best alternative. In the present study, the set

Criteria

Alternatives

Classification Models	Reliability				Time Complexity
	Relationship of parameter	Behaviour of parameter	Matrix of Parameter	Error Rate	
Random Tree	RV (M1/ TS)	MPV (M1/ TS)	BPV (M1/ TS)	ERV (M1/ TS)	TcV (M1/ TS)
Rule. Zero	RV (M2/ TS)	MPV (M2/ TS)	BPV (M2/ TS)	ERV (M2/ TS)	TcV (M2/ TS)
Bayes Net	RV (M3/ TS)	MPV (M3/ TS)	BPV (M3/ TS)	ERV (M3/ TS)	TcV (M3/ TS)
Bayes.NaiveByesUpdateable	RV (M4/ TS)	MPV (M4/ TS)	BPV (M4/ TS)	ERV (M4/ TS)	TcV (M4/ TS)
Lazy.IBK	RV (M5/ TS)	MPV (M5/ TS)	BPV (M5/ TS)	ERV (M5/ TS)	TcV (M5/ TS)
Meta.AdaboostM1	RV (M6/ TS)	MPV (M6/ TS)	BPV (M6/ TS)	ERV (M6/ TS)	TcV (M6/ TS)
Meta.Bagging	…	…	…	…	…
Meta.filteredclassifier	…	…	…	…	…
Meta.logitboost	…	…	…	…	…
Tree.j48	…	…	…	…	…
REPTree	…	…	…	…	…
Random Forest	…	…	…	…	…
Rule. Decision Table:	…	…	…	…	…
Rules. Part	…	…	…	…	…
Meta.RandomCommittee	…	…	…	…	…
Trees.LMT	…	…	…	…	…
Treed.HoeffdingTree	…	…	…	…	…
Kstar	…	…	…	…	…
Functions.smo	…	…	…	…	…
Functions.SIMPLE. logistic	…	…	…	…	…
Byes.NaiveBayes	…	…	…	…	…
Decision Stump	RV (Mn/ TS)	MPV (Mn/ TS)	BPV (Mn/ TS)	ER (Mn/ TS)	TcV (Mn/ TS)

RV: Relationship of parameter Values…….. TcV: Time complexity Values
MPV: Matrix of Parameter Values……. M: Classification model
BPV: Behaviour of parameter values……. TS: Test Samples
ERV: Error Rate Value ……..n: number of Classification models

Fig. 2 Structure of decision matrix

of criteria is obtained from the conducted analysis in the literature.

Step 2. Determination of the best and worst criteria

Considering the best criterion as the most desirable or most important decision criteria is possible, and the worst criterion represents the less desirable or important criteria to the decision. This step involves the description of the best and the worst criteria depending on the perspective of the three decision makers/evaluators. Appendix 1 Section 2 presents the BWM comparison questions and the list of experts.

Step 3. Conduct the pairwise comparison between the best criterion and the other criteria

The pairwise comparison process occurs between the identified best criterion and the other criteria. The aim of this step is to determine the best criterion preference over all the other criteria. The value must be determined by an evaluator/expert and must be from 1 to 9 to represent the importance of the best criterion over the other criteria. This step will result in a vector identified as 'Best-to-Others', which is

$$AB = (a_{B1}, a_{B2}, \ldots, a_{Bn}),$$

where a_{Bj} indicates the importance of the best criterion$B$ over criterion j, and $a_{BB} = 1$.

Step 4. Pairwise comparison process between the other criteria and the worst criterion

The aim of comparison is to identify the preference for all the criteria over the least important criterion. The importance is determined by an evaluator/expert of all the criteria over the worst criterion, and the numbers from 1 to 9 are used towards indicating the importance. The result for this step is a vector recognised as 'Others-to-Worst'. The vector result of 'Others-to-Worst' is represented as $A_w = (a_{1w}, a_{2w}, \ldots, a_{aw})$, where a_{jw}

represents the preference of the criterion j over the worst criterion W. Clearly, $a_{ww} = 1$. Two types of reference comparisons, namely, Best-to-Others and Others-to-Worst criteria, are illustrated in Fig. 3.

Step 5. Elicit the optimal weights (W^*1, W^*2, …W^*n)

The optimal weight for the criteria is the one where for each pair of W_B/Wj and Wj/Ww, $W_B/W_j = a_{BJ}$ and $W_j / W_w = a_{jw}$.

To fulfil these conditions for all j, a solution where the maximum absolute differences for all j are minimised must be obtained:

$$\left|\frac{W_B}{W_j} - a_{Bj}\right| and \left|\frac{W_i}{W_w} - a_{Bjw}\right| \tag{4}$$

Considering the non-negativity and sum condition for the weights, the following problem is created:

$$\min\max_j \left\{\left|\frac{W_B}{W_j} - a_{Bj}\right|, \left|\frac{W_j}{W_w} - a_{jw}\right|\right\} \tag{5}$$

$$W_j \geq 0, \text{ for all } j$$
$$\sum_j W_{j=1}$$

The aforementioned problem can be transferred to the following problem:

$$\min \xi,$$
s.t.

$$\left|\frac{W_B}{W_J} - a_{Bj}\right| \leq \xi, for\ all\ j \tag{6}$$

$$\left|\frac{W_j}{W_w} - a_{jw}\right| \leq \xi, for\ all\ j \tag{7}$$

$$\sum_j W_{j=1}$$
$$W_j \geq 0, \text{ for all } j$$

By finding a solution for the last problem, the optimal weights (w^*_1; w^*_2;…; w^*_n) and ξ, n are obtained. The value for ξ^* reflects the outcomes' reliability, depending on the extent of consistency in the comparisons. A value close to zero represents high consistency, and thus, high reliability [112, 126, 127, 148]. After that, the ratio for consistency calculated by using ξ^* and the corresponding consistency index is as follows (Table 1):

$$\text{Consistency Ratio} = \frac{\xi^*}{Consistency\ Index} \tag{8}$$

As proposed by [112], the bigger the ξ^* is, the more consistent the vectors are.

Ranking the multiclass classification models based on VIKOR method Owing to the suitability of VIKOR for many alternatives and multiple conflicting criteria decision cases, it is used to rank multiclass classification models. VIKOR can provide rapid results, thereby determining the most suitable option at the same time. The weights for all the criteria will be gathered from the BWM and will be utilised in VIKOR. The results for the decision alternative are ranked in ascending order. The models of multiclass classification are ranked based on values of weighted criteria that employ the VIKOR method. VIKOR steps are presented below [149, 150].

Step 1: Identify the best f^*i and worst f^-i values of all criterion functions, i = 1; 2; ...; n. If the ith function represents a benefit, then

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

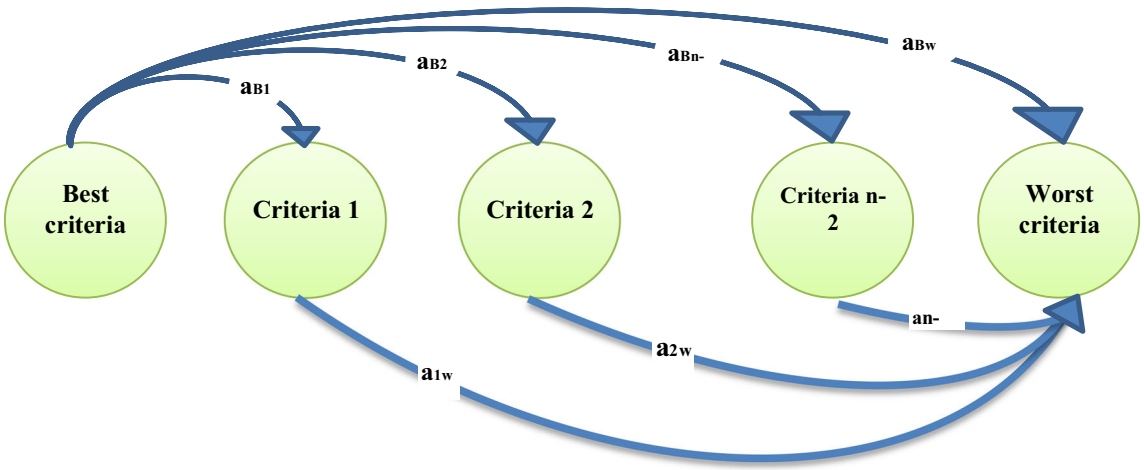

Fig. 3 Reference comparisons in the BWM method

Table 1 Index of consistency

a_{BW}	1	2	3	4	5	6	7	8	9
Consistency index	0.0	0.44	1.0	1.63	2.30	3.00	3.73	4.47	5.23

Step 2:

Based on the BWM method, the weights for each criterion are computed. A set of weights $w = w_1, w_2, w_3, \cdots, w_j, \cdots, w_n$ from the decision maker is accommodated in the DM. This set is equal to 1. The resulting matrix can also be computed as demonstrated in following equation.

$$WM = wi * \frac{f^{*}i - fij}{f^{*}i - f^{-}i} \tag{10}$$

This process will produce a weighted matrix as follows:

$$\begin{bmatrix} w_1\left(f^{*}1 - f11\right)/\left(f^{*}1 - f^{-}1\right) & w_2\left(f^{*}2 - f12\right)/\left(f^{*}2 - f^{-}2\right) & \cdots & w_1\left(f^{*}\mathrm{i} - fij\right)/\left(f^{*}\mathrm{i} - f^{-}\mathrm{i}\right) \\ w_1\left(f^{*}1 - f21\right)/\left(f^{*}1 - f^{-}1\right) & w_2\left(f^{*}2 - f22\right)/\left(f^{*}2 - f^{-}2\right) & \cdots & w_1\left(f^{*}\mathrm{i} - fij\right)/\left(f^{*}\mathrm{i} - f^{-}\mathrm{i}\right) \\ \vdots & \vdots & \vdots & \vdots \\ w_1\left(f^{*}1 - f31\right)/\left(f^{*}1 - f^{-}1\right) & w_2\left(f^{*}2 - f32\right)/\left(f^{*}2 - f^{-}2\right) & \cdots & w_1\left(f^{*}\mathrm{i} - fij\right)/\left(f^{*}\mathrm{i} - f^{-}\mathrm{i}\right) \end{bmatrix} \tag{11}$$

Step 3:

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

$$Sj = \sum_{i=1}^{n} wi * \frac{f^{*}i - fij}{f^{*}i - f^{-}i} \tag{12}$$

$$Rj = \max_{i} wi * \frac{f^{*}i - fij}{f^{*}i - f^{-}i} \tag{13}$$

where *wi* indicates the criterion weights expressing their relative importance.

Step 4:

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

$$Q_{\mathrm{j}} = \frac{\mathrm{v}\left(S_{\mathrm{j}} - S^{*}\right)}{S^{-} - S^{*}} + \frac{(1 - \mathrm{v})\left(R_{\mathrm{j}} - R^{*}\right)}{R^{-} - R^{*}} \tag{14}$$

where

$$S^{*} = \min_{j} S_j, S^{-} = \max_{j} S_j \quad R^{*} = \min_{j} R_j R^{-} = \max_{j} R_j$$

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

Step 5:

The alternatives can now be ranked by sorting the values of S, R and Q in ascending order. Optimal performance is indicated by the lowest value.

Step 6:

Propose as a compromise solution alternative (a'), which ranks best by the measure Q (minimum) if the following two conditions are satisfied:

C1. 'Acceptable advantage':

$$\mathrm{Q}\left(a''\right) - \mathrm{Q}\left(a'\right) \geq \mathrm{DQ} \tag{15}$$

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

C2. 'Stability' is acceptable in the decision-making context. Alternative a' should also be ranked best by S and/or R. This compromise solution is stable within the process of decision making, which can be 'voting by majority rule' (v > 0:5), 'by consensus' (v ≅0.5) or 'with veto' (v < 0.5). Here, v is the decision-making strategy weight of 'the majority of criteria' (or 'the maximum group utility'). The Q value provides an idea of which multiclass classification model has higher values of evaluation criteria than the others. According to this technique, the multiclass classification models with high values of evaluation criteria will have the lowest Q value. Two main decision-making contexts will be applied: individual decision making and GDM. In the former, decision making will be based on a single individual decision maker, whereas GDM is based on multiple decision makers/experts. GDM will be performed in two ways: internal aggregation and external

aggregation. Figure 4 illustrates the procedures that will be followed to apply the types of aggregation.

Figure 4 shows that the internal GDM is calculated by using the arithmetic mean of the final weights of the three experts' preferences to eliminate the possible variation among them. VIKOR is then applied based on final weights obtained from the arithmetic mean of the three experts. By contrast, external aggregation is calculated by using the arithmetic mean of the Q values for each expert's ranking, and then the final Q values depend on external group ranking.

Results and discussion

This section presents the results of the proposed framework of evaluation and benchmarking the multiclass classification models of acute leukaemia. Section 5.1 presents the data in decision matrix. Section 5.2 presents the results of the development in benchmarking framework that involves BWM results in subsection 5.2.1 to show the weights for the main criteria and subs-criteria and the results of the VIKOR method in subsection 5.2.2. Section 5.3 presents the validation processes and results.

Data presentation in decision matrix

The results obtained from the evaluation of the 22 multiclass classification models are presented in this section. The outcome of the implementation process of those 22 multiclass classification models generated four parameters (tp, tn, fp, fn) which are considered fundamental values to calculate the rest reliability criteria group values. The values of time complexity criterion were calculated according to its respective framework. The values of reliability group of criteria and time complexity criterion were considered an input to fill the decision matrix. Table 2 illustrates the completed decision matrix.

Table 2 shows that each multiclass classification model has been evaluated based on 11 evaluation criteria. The next section will discuss in detail the results of integration between the BWM and VIKOR.

Results of the framework of evaluation and benchmarking multiclass classification models

The results of the proposed benchmarking framework are represented in two subsections. The first section is the weight result by using the BWM, whereas the second is the result of using VIKOR. The VIKOR section is divided into the individual context and the group context. The group context includes the result of the internal and external aggregation, which will be described in detail in subsequent sections.

Results for weight using BWM method

In this section, BWM results are presented and explained. Three experts were asked to make their evaluation and benchmarking preferences on criteria of multiclass classification models via BWM comparison questions. Table 3 presents the first expert's process results of main criteria and their subcriteria. Appendix 2 (Tables 21 and 22) shows the detailed results of the other two experts.

R: Reliability, TM: Time Complexity, MOP: Matrix of parameter, ROP: Relationship of parameter, BOP: Behaviour of parameter, ER: Error Rate, True Positive: TP, True Negative: TN, FP: False Positive, FN: False Negative. Table 3 and

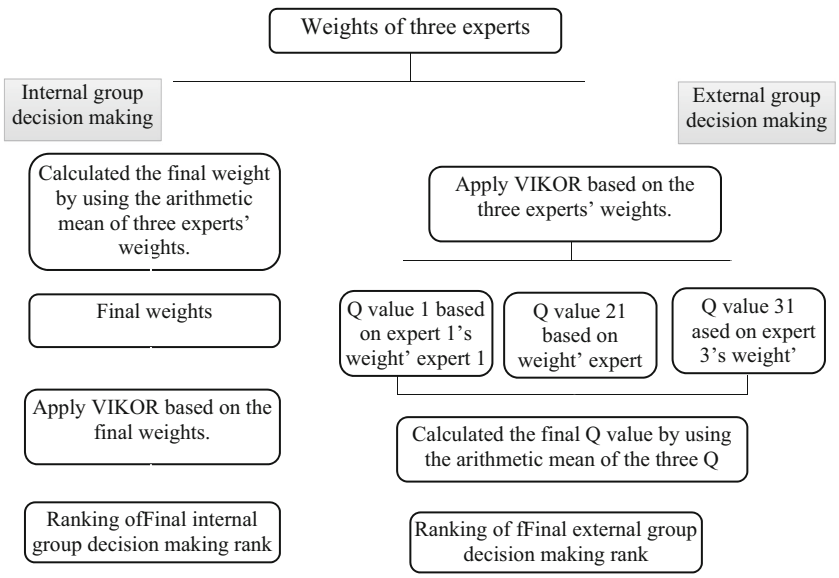

Fig. 4 Internal and external aggregation

Table 2 The decision matrix

	Models	Ave-accuracy	$Precision_{\mu}$	$Precision_{M}$	$Recall_{M}$	F-$score_{M}$	Error Rate	TP	TN	FP	FN	Time complexity
1	Rule.Zero	0.694	0.541	0.843	0.379	0.391	0.694	13.333	38.000	11.333	11.333	0.1
2	BayesNet	0.973	0.959	0.941	0.604	0.878	0.973	23.667	47.333	1.000	1.000	0.53
3	Bayes.NaiveByesUpdateable	0.973	0.959	0.976	0.579	0.879	0.973	23.333	47.667	1.000	1.000	0.14
4	Lazy.IBK	0.748	0.622	0.509	0.268	0.106	0.748	15.333	40.000	9.333	9.333	0.1
5	Meta.AdaboostM1	0.845	0.767	0.843	0.345	0.346	0.845	18.667	43.000	5.667	5.667	1.06
6	Meta.Bagging	0.799	0.699	0.804	0.328	0.299	0.799	17.000	41.333	7.333	7.333	0.94
7	Meta.filteredclassifier	0.662	0.493	0.425	0.195	0.051	0.662	12.000	36.333	12.333	12.333	0.48
8	Meta.logitboost	0.790	0.685	0.552	0.294	0.137	0.790	16.667	41.000	7.667	7.667	1.11
9	Tree.j48	0.616	0.425	0.368	0.187	0.038	0.616	10.333	34.667	14.000	14.000	0.45
10	REPTree	0.799	0.699	0.810	0.337	0.312	0.799	17.000	41.333	7.333	7.333	0.22
11	RandomTree	0.587	0.381	0.367	0.207	0.044	0.587	8.000	29.000	13.000	13.000	0.05
12	RandomForest	0.854	0.781	0.868	0.354	0.375	0.854	19.000	43.333	5.333	5.333	0.44
13	Rule. Decision Table:	0.781	0.671	0.558	0.328	0.162	0.781	16.333	40.667	8.000	8.000	3.19
14	Rules.part	0.607	0.411	0.344	0.183	0.033	0.607	10.000	34.333	14.333	14.333	0.20
15	Meta.RandomCommittee	0.790	0.685	0.641	0.328	0.204	0.790	16.667	41.000	7.667	7.667	0.05
16	Trees.LMT	0.763	0.644	0.539	0.303	0.137	0.763	15.667	40.000	8.667	8.667	3.16
17	Treed.HoeffdingTree	0.671	0.507	0.446	0.216	0.064	0.671	12.333	36.667	12.000	12.000	0.68
18	Kstar	0.698	0.547	0.845	0.379	0.392	0.698	13.667	38.667	11.333	11.333	0.36
19	Functions.Smo	0.793	0.689	0.557	0.285	0.134	0.793	17.000	41.667	7.667	7.667	0.49
20	Functions.SIMPLE.logistic	0.763	0.644	0.539	0.303	0.137	0.493	15.667	40.000	8.667	8.667	1.07
21	Byes.NaiveBayes	0.644	0.466	0.420	0.165	0.041	0.547	11.333	35.667	13.000	13.000	0.08
22	Decision Stump	0.845	0.767	0.840	0.345	0.343	0.845	18.667	43.000	5.667	5.667	0.09

Appendix 2 (Tables 21 and 22) present the three experts' processes weighted results based on BWM. For the evaluation and benchmarking criteria, the best and worst criteria are identified, the best criteria is compared with the other criteria, and the worst criterion is determined. Lastly, the linear model of BWM solved according to Eqs. (6, 7) in Sect. 4.2.1.1 to obtain the weights. Eq. (8) has been used to calculate the consistency ratio of each expert's preferences. To calculate the global weights of each criterion for the three experts, BWM method derives the local weights for each criteria group at each level as shown in Table 3 and Appendix 2 (Tables 21 and 22) that explains the importance of each criterion regarding the parent. Consequently, the global weights for each criterion is obtained. Each global weight explains each criterion's importance with respect to the goal for each expert. Firstly, the weight of each criterion was determined by making a comparison between criteria based on BWM. These weights are called 'local weights'. To find the global weights with respect to the goal, the criteria's origin weights and their associated local weights were multiplied, as presented in Table 4.

Table 4 presents the overall local and global weights for the three *e*xperts for 11 evaluation and benchmarking criteria. The overall CR for the three experts scores an acceptable ratio of less than 0.1. These global weights have been used in our benchmarking framework because the global weights represent the importance of the criteria with respect to the goal. Table 4 shows that the global weight results of the first expert assigned the maximum weight for true positive with a value of 0.201. The minimum weight obtained by $precision_{M}$ and $recall_{M}$ is 0.035 and 0.035, respectively. The second expert assigned the maximum weight for time complexity criterion with a value of 0.500. The minimum weight obtained by ave-accuracy is 0.011. The third expert assigned the maximum weight for time complexity with a value of 0.200. The minimum weight obtained by true negative is 0.015. Final weight results are used in applying VIKOR method the next section.

Ranking's results of VIKOR method

The results after the ranking of the multiclass classification models based on weighted evaluation criteria are presented in this section. Individual decision making and GDM contexts are explained. The results of the individual and group VIKOR decision-making contexts are presented in the following subsections.

Table 3 Results of the BWM method for weight preferences of the criteria of evaluation and benchmarking the multiclass classification (first expert)

Expert 1							
Level 1 of Criteria: Main Criteria							
List of criteria	Best criterion	Other Criteria	Scores				Weight
R	$_R$	TC	5				0.833
TC		–					0.167
Consistency: 0							
Level 2 of Criteria: sub criteria of Reliability							
List of criteria	Best criterion	Other Criteria	Scores	Other criteria	Worst criterion	Scores	Weight
MOP	MOP	ROP	3	ROP	ER	3	0.577
ROP		BOP	5	BOP		2	0.210
BOP		ER	6	MOP		6	0.126
ER		–	–	–		–	0.087
Consistency: 0.017							
Level 3 of Criteria: sub criteria of Matrix of parameter							
List of criteria	Best criterion	Other Criteria	Scores	Other criteria	Worst criterion	Scores	Weight
TP	TP	TN	3	FP	TN	2	0.419
TN		FP	2	FN		2	0.129
FP		FN	2	TP		3	0.226
FN		–		–			0.226
Consistency: 0.032							
Level 3 of Criteria: sub criteria of Relationship of parameter							
List of criteria	Best criterion	Other Criteria	Scores	Other criteria	Worst criterion	Scores	Weight
Ave Accuracy	$Precision_{\mu}$	Ave Accuracy	5	$Precision_{\mu}$	Ave Accuracy	5	0.079
$Precision_{\mu}$		$Precision_M$	2	$Precision_M$		3	0.487
$Precision_M$		$Recall_M$	4	$Recall_M$		3	0.289
$Recall_M$		–		–		–	0.145
Consistency: 0.040							

- VIKOR Results of Individual Context for Different *Experts'* Weights

VIKOR is utilised to rank alternatives based on the decision matrix results presented in Table 2 and the results of the weights presented in Table 4. The ranking show the importance of the evaluation criteria from the viewpoint of each *expert*. VIKOR technique depends on Q value in ranking the alternatives. The alternative with a lower Q value is considered the better alternative, whereas the alternative with a higher Q value is considered the worst alternative. Table 5 shows the VIKOR results of ranking according to the weights that reflect the viewpoint of the first expert. Tables 23 and 24 in Appendix 3 show the VIKOR results of the two other experts.

Table 5 and Appendix 3 (Tables 23 and 24) present the three VIKOR ranking results provided by the experts. In the first rank, 'Bayes.NaiveByesUpdateable' had the lowest Q value of 0.0358 for and was thus the best multiclass classification model in this rank. By contrast, 'RandomTree' had the highest Q value of 1 and was thus the worst multiclass classification model in this rank. In the second rank, 'Byes.NaiveBayes' had the lowest Q value of 0 and was thus the best multiclass classification model in this rank. By contrast, 'Rule.Decision Table' had the highest Q value 1 and was thus the worst multiclass classification model. In the third rank, the lower Q value was 0 for 'Bayes.NaiveByesUpdateable', which was eventually considered the best multiclass classification model in this rank. By contrast, the higher Q value was 0.9956 for 'Rules.part', which was considered the worst. Differences in weight provided by the experts affected the ranking scores. Figure 5 shows the variance among the VIKOR results.

Figure 5 demonstrates the final VIKOR ranking for three experts. Ten classification models were selected from each score ranking results [2]. The selected classification models with the best score received the highest ranking (first five classification models), whereas the classification models with the worst score received the lowest ranking (last five classification models).

The first five classification models with the highest-ranking level vary with regard to the weights provided by the experts. According to the weights provided by expert

Table 4 BWM local and global weights for three expert

First Expert							
Main criteria	Weights	Sub-criteria	Local Weights	Global Weights	Sub-criteria	Local Weights	Global Weights
Reliability	0.833	MOP	0.577	0.481	TP	0.419	0.201
					TN	0.129	0.062
					FP	0.226	0.109
					FN	0.226	0.109
		ROP	0.210	0.175	Ave Accuracy	0.079	0.014
					$Precision_{\mu}$	0.487	0.085
					$Precision_{M}$	0.289	0.051
					$Recall_{M}$	0.145	0.025
		BOP	0.126				0.105
		ER	0.087				0.072
Time Complexity							0.167
Overall Consistency Ratio > 0.1							
Third Expert							
Reliability	0.800	MOP	0.149	0.119	TP	0.305	0.037
					TN	0.130	0.015
					FP	0.217	0.026
					FN	0.348	0.041
		ROP	0.537	0.430	Ave Accuracy	0.154	0.066
					$Precision_{\mu}$	0.359	0.154
					$Precision_{M}$	0.231	0.099
					$Recall_{M}$	0.256	0.110
		BOP	0.249				0.199
		ER	0.065				0.052
Time Complexity							0.200
Overall Consistency Ratio > 0.1							

Second Expert							
Main criteria	Weights	Sub-criteria	Local Weights	Global Weights	Sub-criteria	Local Weights	Global Weights
Reliability	0.500	MOP	0.147	0.074	TP	0.562	0.042
					TN	0.237	0.017
					FP	0.142	0.010
					FN	0.059	0.004
		ROP	0.244	0.122	Ave Accuracy	0.093	0.011
					$Precision_{\mu}$	0.569	0.069
					$Precision_{M}$	0.169	0.021
					$Recall_{M}$	0.169	0.021
		BOP	0.054				0.027
		ER	0.555				0.278
Time Complexity							0.500
Overall Consistency Ratio > 0.1							

Table 5 Ranking results based on the first expert's weights

Machine learning	Q	Order
Rule.zero	0.5836	14
BayesNet	0.0501	2
Bayes.NaiveByesUpdateable	0.0358	1
Lazy.IBK	0.4894	12
Meta.AdaboostM1	0.2419	5
Meta.Bagging	0.3594	7
Meta.filteredclassifier	0.7572	16
Meta.logitboost	0.4295	10
Tree.j48	0.8825	20
REPTree	0.3286	6
RandomTree	1	22
RandomForest	0.1906	3
Rule. Decision Table:	0.7875	18
Rules.part	0.8988	21
Meta.RandomCommittee	0.3668	8
Trees.LMT	0.8029	19
Treed.HoeffdingTree	0.7375	15
Kstar	0.5731	13
Functions.Smo	0.4016	9
Functions.SIMPLE.logistic	0.4634	11
Byes.NaiveBayes	0.7814	17
Decision Stump	0.2048	4

one (A) and expert three (C), Bayes.NaiveByesUpdateable and BayesNet models appeared in the first and second indices, respectively. By contrast, the first and second indices based on the weights provided by expert two (B) were Byes.NaiveBayes and RandomTree. Random Forest and Decision Stump appeared in the third and fourth indices based on the weight provided by expert (A) and expert (C,) whereas the two classification models did not appear in first five indices according to the second expert. Rules.part and Rule.zero were in the third and fourth indices based on the weight provided by expert (B). Meta.AdaboostM1 was in the fifth index according to the weight given by expert (A), whereas Rule.zero appeared in the fifth index based on the weigh obtained from expert (B) and expert (C).

The last five classification models considered with the lowest-ranking level vary based on the weights provided by the expert. Accordingly, RandomTree is the worst model with index 22 according to expert (A), whereas the same model was in the third worst classification model based on expert (C). The worst one according to expert (B) is Rule. Decision Table, in additional the same model was the fifth worst model according to experts (A) and (C). Rules.part appeared as the worst classification model based on expert (C) and the second worst classification model according to expert (A). Trees.LMT was the second worst classification model according to expert (B). In the same last classification model, it was the fourth worst classification model according to experts (A) and (C). Tree.j48 is the third worst model according to expert (A) and the second worst model according to expert (C). Lastly, Meta.AdaboostM1 and Meta.logitboost were the fourth and fifth worst classification models, respectively, based on expert (B).

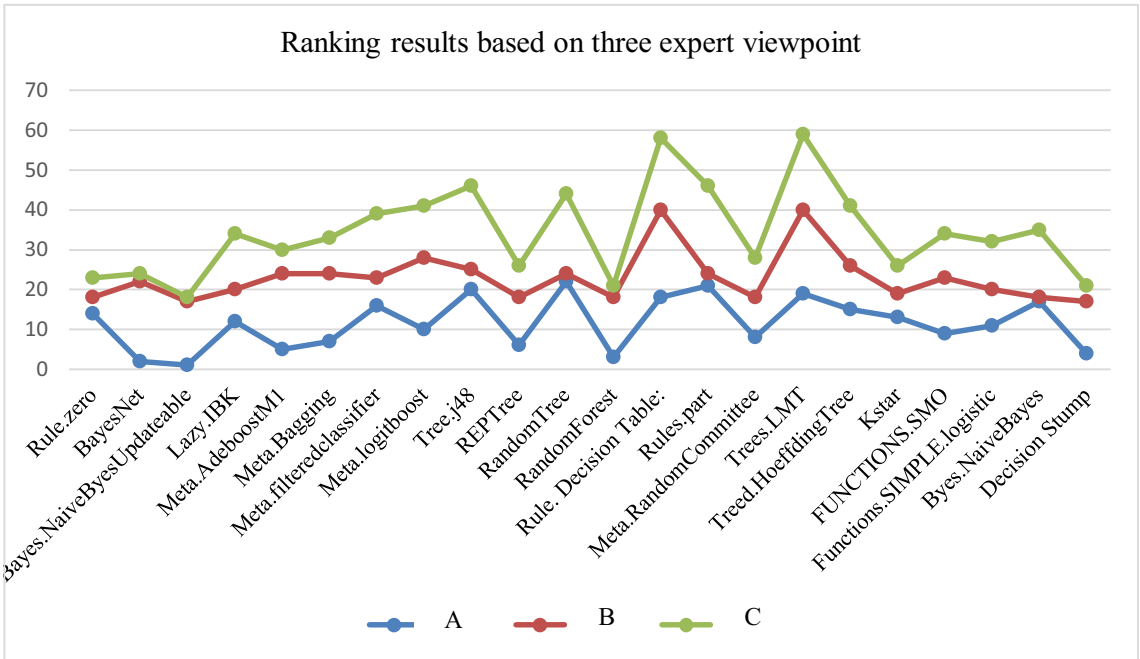

Fig. 5 Ranking results based on the three experts' weights. (A) First expert's ranking, (B) second expert's ranking, (C) third expert's ranking

The results of the individual context clearly show variances among the rankings of three experts. Therefore, the utilisation of group VIKOR decision-making context, which aims to provide ranking alternatives which in turn considers overall decision makers, is necessary. The following sections present the results of group VIKOR decision-making context.

- Group VIKOR with Internal and External Aggregation

To extend VIKOR into a group decision environment, two ways were used; (1) internal and (2) external aggregation, both of which depend on multiple decision makers. Internal GDM results are calculated by using the arithmetic mean of the final weighs of the three experts' preferences to eliminate the variance between them, then the VIKOR is applied based on final arithmetic mean results. By contrast, external aggregation results are calculated by finding the arithmetic mean of the Q values for each expert's ranking results. The final Q values then depend on the external group ranking. Table 6 illustrates the overall ranking results of VIKOR with internal and external group decision making for 22 multiclass classification models.

As shown in Table 6, the order of the best/first three classification models are Bayes.NaiveByesUpdateable, BayesNet and Decision Stump. The order of the last worst/two classification models based on the results of internal and external GDM are Trees.LMT and Rule. Decision Table. The rest of the classification models with the same order in both internal and external decision making are Meta.RandomCommittee, Lazy.IBK, Meta.logitboost and Byes.NaiveBayes in the following order 8, 13, 14, 15, respectively. By contrast, some classification models are ranked differently between the internal and external group decision making. The order of those classification models based on internal ranking are as follows: REPTree, Rule.zero, RandomForest, Kstar, Meta.Bagging, Meta.AdaboostM1, Functions.SIMPLE. logistic, Functions.Smo, Meta.filteredclassifier, Treed.HoeffdingTree, RandomTree, Rules.part and Tree.j48 in the following order: 4, 5, 6, 7, 9, 10, 11, 12, 16, 17, 18 and 19, respectively. The order of the same classification models based on external ranking aere REPTree, Rule.zero, RandomForest, Kstar, Meta.Bagging, Meta.AdaboostM1, Functions.SIMPLE. logistic, Functions.Smo, Meta.filteredclassifier, Treed.HoeffdingTree, RandomTree, Rules.part and

Table 6 Overall ranking results of VIKOR with internal and external group decision making

Internal group decision making			External group decision making		
Machine learning	Q	Order	Machine learning	Q	Order
Rule.zero	0.2342	5	Rule.zero	0.3946	7
BayesNet	0.1706	2	BayesNet	0.1429	2
Bayes.NaiveByesUpdateable	0.1404	1	Bayes.NaiveByesUpdateable	0.1142	1
Lazy.IBK	0.3356	13	Lazy.IBK	0.4849	13
Meta.AdaboostM1	0.2941	10	Meta.AdaboostM1	0.3701	6
Meta.Bagging	0.2895	9	Meta.Bagging	0.4168	10
Meta.filteredclassifier	0.4497	16	Meta.filteredclassifier	0.6186	17
Meta.logitboost	0.3774	14	Meta.logitboost	0.5174	14
Tree.j48	0.4826	20	Tree.j48	0.6625	19
REPTree	0.2292	4	REPTree	0.3564	5
RandomTree	0.4619	18	RandomTree	0.6787	20
RandomForest	0.2395	6	RandomForest	0.3038	4
Rule. Decision Table	0.9856	21	Rule. Decision Table	0.9155	21
Rules.part	0.4718	19	Rules.part	0.6547	18
Meta.RandomCommittee	0.2531	8	Meta.RandomCommittee	0.4022	8
Trees.LMT	0.9936	22	Trees.LMT	0.9217	22
Treed.HoeffdingTree	0.4517	17	Treed.HoeffdingTree	0.6186	16
Kstar	0.2492	7	Kstar	0.4065	9
Functions.Smo	0.3252	12	Functions.Smo	0.4708	11
Functions.SIMPLE. logistic	0.3222	11	Functions.SIMPLE. logistic	0.4736	12
Byes.NaiveBayes	0.4121	15	Byes.NaiveBayes	0.5768	15
Decision Stump	0.2173	3	Decision Stump	0.2985	3

Tree.j48 in the following order: 5, 7, 4, 9, 10, 6, 12, 11, 17, 16, 20, 18 and 19, respectively. Therefore, the first best three index classification models in both internal and external GDM are equal, whereas the last worst two index classification models are equal as well. The fourth classification models in different medium scores indices were equal, whereas the rest of the classification models showed different score indices. From this point forward, the internal and external aggregation decision making rank can be considered the final ranking results and will be used in validation processes. The next section will describe in detail the validation results.

Validation processes and results

Decision selection of multiclass classification model is considered a difficult task because it relies on conflicting multiple criteria in one side. Differences in accuracy, performance and other features make the task difficult. The results are validated for the proposed benchmarking framework by utilising objective validations.

Objective validation

Statistical methods of mean and standard deviation (SD) were used in this study to ensure that multiclass classification models were ranked according to the proposed benchmarking framework. Towards this goal, three groups were created and separated because of the results ranking for multiclass classification models [2, 82]. Each group's results are expressed as mean ± SD. The mean is the average results. Its calculation is performed by the sum division of the observed results over the resulting number and by the following equation:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{16}$$

SD is used to determine the dispersion or variation amount in the set of values and is calculated by the following equation:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2} \tag{17}$$

The utilisation of mean ± SD ensures that the three multiclass classification models sets are subject to systematic ordering. The multiclass classification models scoring was divided into three groups to validate the results ranking by using the above test. Division took place based on the ranking result obtained from the proposed benchmarking framework. An equal number (seven) are included for the first and second multiclass classification models. Eight classification models (2) are included in the third group depending on the scoring values from the ranking results. For this process to takes place, two statistical methods will be used. These methods must prove that the lower scoring value was achieved by the first group when both mean and SD are measured. The lower mean and SD were assumed for the first group in comparison with the other two groups to validate the results. The results for both mean and SD of the second group must be lower or equal than the ones in the third group. At the same time, they must be higher than the first group. Nevertheless, results of the mean and SD must be higher than those in the first and second groups and equal to those in the second group. The results of the first group must be statistically proven according to the systematic ranking results which have to be considered lowest among the three groups.

Validation results

This section presents the validation processes of internal and external GDM ranking. In this research, objective validation processes are used. The validation process for multiclass classification models ranking results has been obtained by dividing the ranking result into three groups. The first two groups are equal, with each one having 7 models and the third one having 8 models. The mean ± SD have been calculated for each group to ensure that the ranking multiclass classification models undergo a systematic ranking. After normalisation and weighting process for the row data of the first, second and third groups of multiclass classification models, the validation results for internal and external GDM are presented in Table 7.

Table 7 shows the results of validation for internal aggregation group decision making. The first group has a lower mean ± SD than the second group except for error rate (M = 0.0951 ± 0.0319 in the first group; M = 0.0721 ± 0.0327 in the second group). For the second group; the mean ± SD is lower than the mean ± SD in third group for all features except for error rate (M = 0.0721 ± 0.0327 in the second group; M = 0.0450 ± 0.0231 in the third group). Accordingly, first group has a lower value compared with the second group. The second group has a lower value compared with the third group. Regarding the results of validation for external aggregation GDM, the mean ± SD in the first group is lower than the mean ± SD in the second group except for error rate (M = 0.1010 ± 0.0272 in the first group; M = 0.0662 ± 0.0309 in the second group). In the second group, the mean ± SD is lower than the mean ± SD in the third group for all features except for error rate (M = 0.0662 ± 0.0309 in the second group; M = 0.0450 ± 0.0231 in the third group). Accordingly, the first group has a lower value compared with the second group, whereas the second group has a lower value compared with the third group.

Table 7 Validation results of internal and external group decision making rank

Machine learning	Time	Macro-precision	Micro-Precision	Macro-recall	F-score	Ave-accuracy	Error Rate	TP	FP	TN	FN
Internal aggregation for group decision making											
Bayes.NaiveByesUpdateable	0.0083	0	0.0001	0.0029	0	0	0.1340	0.0020	0	0	0
BayesNet	0.0442	0.0031	0	0	0.0002	0	0.1340	0	0	0.0006	0
Decision Stump	0.0037	0.0122	0.0342	0.0306	0.0699	0.0101	0.0983	0.0297	0.0169	0.0079	0.0180
REPTree	0.0156	0.0149	0.0464	0.0316	0.0739	0.0137	0.0855	0.0396	0.0229	0.0107	0.0245
Rule.zero	0.0046	0.0120	0.0745	0.0266	0.0637	0.0220	0.0560	0.0614	0.0374	0.0164	0.0399
RandomForest	0.0359	0.0097	0.0318	0.0296	0.0657	0.0093	0.1008	0.0277	0.0157	0.0073	0.0167
Kstar	0.0285	0.0118	0.0735	0.0266	0.0635	0.0216	0.0572	0.0594	0.0374	0.0153	0.0399
Mean	0.0201	0.0091	0.0372	0.0211	0.0481	0.0110	0.0951	0.0314	0.0186	0.0083	0.0199
Std	0.0162	0.0054	0.0305	0.0136	0.0330	0.0090	0.0319	0.0245	0.0154	0.0064	0.0165
Overall mean	0.0291										
Overall std	0.0184										
Meta.RandomCommittee	0	0.0301	0.0489	0.0327	0.0881	0.0144	0.0829	0.0416	0.0242	0.0113	0.0257
Meta.Bagging	0.0819	0.0154	0.0464	0.0327	0.0757	0.0137	0.0855	0.0396	0.0229	0.0107	0.0245
Meta.AdaboostM1	0.0930	0.0119	0.0342	0.0306	0.0695	0.0101	0.0983	0.0297	0.0169	0.0079	0.0180
Functions.SIMPLE.logistic	0.0939	0.0393	0.0562	0.0357	0.0968	0.0165	0	0.0475	0.0278	0.0130	0.0296
Functions.Smo	0.0405	0.0377	0.0481	0.0377	0.0972	0.0142	0.0837	0.0396	0.0242	0.0102	0.0257
Lazy.IBK	0.0046	0.0420	0.0601	0.0397	0.1009	0.0177	0.0711	0.0495	0.0302	0.0130	0.0322
Meta.logitboost	0.0976	0.0381	0.0489	0.0367	0.0968	0.0144	0.0829	0.0416	0.0242	0.0113	0.0257
Mean	0.0588	0.0306	0.0490	0.0351	0.0893	0.0144	0.0721	0.0413	0.0243	0.0111	0.0259
Std	0.0431	0.0122	0.0082	0.0032	0.0121	0.0024	0.0327	0.0064	0.0042	0.0018	0.0044
Overall mean	0.0411										
Overall std	0.0119										
Byes.NaiveBayes	0.0028	0.0500	0.0879	0.0520	0.1094	0.0259	0.0151	0.0732	0.0435	0.0203	0.0463
Meta.filteredclassifier	0.0396	0.0496	0.0830	0.0484	0.1080	0.0244	0.0472	0.0693	0.0411	0.0192	0.0438
Treed.HoeffdingTree	0.0580	0.0476	0.0805	0.0459	0.1064	0.0237	0.0498	0.0673	0.0399	0.0186	0.0425
RandomTree	0.0000	0.0548	0.1029	0.0470	0.1090	0.0303	0.0263	0.0930	0.0435	0.0316	0.0463
Rules. Part	0.0138	0.0568	0.0976	0.0499	0.1104	0.0288	0.0319	0.0811	0.0483	0.0226	0.0515
Tree.j48	0.0368	0.0546	0.0952	0.0494	0.1097	0.0280	0.0344	0.0792	0.0471	0.0220	0.0502
Rule. Decision Table:	0.2890	0.0376	0.0513	0.0327	0.0935	0.0151	0.0804	0.0435	0.0254	0.0119	0.0270
Trees.LMT	0.2862	0.0393	0.0562	0.0357	0.0968	0.0165	0.0753	0.0475	0.0278	0.0130	0.0296
Mean	0.0908	0.0488	0.0818	0.0451	0.1054	0.0241	0.0450	0.0693	0.0396	0.0199	0.0422
Std	0.1231	0.0071	0.0189	0.0070	0.0065	0.0056	0.0231	0.0167	0.0085	0.0061	0.0091
Overall mean	0.0556										
Overall std	0.0211										
Validation results for the external aggregation group decision making											
Bayes.NaiveByesUpdateable	0.0083	0	0.0001	0.0029	0	0	0.1340	0.0020	0	0	0
BayesNet	0.0442	0.0031	0	0	0.0002	0	0.1340	0	0	0.0006	0
Decision Stump	0.0037	0.0122	0.0342	0.0306	0.0699	0.0101	0.0983	0.0297	0.0169	0.0079	0.0180
RandomForest	0.0359	0.0097	0.0318	0.0296	0.0657	0.0093	0.1008	0.0277	0.0157	0.0073	0.0167
REPTree	0.0156	0.0149	0.0464	0.0316	0.0739	0.0137	0.0855	0.0396	0.0229	0.0107	0.0245
Meta.AdaboostM1	0.0930	0.0119	0.0342	0.0306	0.0695	0.0101	0.0983	0.0297	0.0169	0.0079	0.0180
Rule.zero	0.0046	0.0120	0.0745	0.0266	0.0637	0.0220	0.0560	0.0614	0.0374	0.0164	0.0399
Mean	0.0293	0.0091	0.0316	0.0217	0.0490	0.0093	0.1010	0.0271	0.0157	0.0073	0.0167
Std	0.0322	0.0055	0.0260	0.0140	0.0336	0.0077	0.0272	0.0212	0.0130	0.0057	0.0139
Overall mean	0.0289										
Overall std	0.0182										
Meta.RandomCommittee	0.0000	0.0301	0.0489	0.0327	0.0881	0.0144	0.0829	0.0416	0.0242	0.0113	0.0257

Table 7 (continued)

Machine learning	Time	Macro-precision	Micro-Precision	Macro-recall	F-score	Ave-accuracy	Error Rate	TP	FP	TN	FN
Kstar	0.0285	0.0118	0.0735	0.0266	0.0635	0.0216	0.0572	0.0594	0.0374	0.0153	0.0399
Meta.Bagging	0.0819	0.0154	0.0464	0.0327	0.0757	0.0137	0.0855	0.0396	0.0229	0.0107	0.0245
Functions.Smo	0.0405	0.0377	0.0481	0.0377	0.0972	0.0142	0.0837	0.0396	0.0242	0.0102	0.0257
Functions.simple.logistic	0.0939	0.0393	0.0562	0.0357	0.0968	0.0165	0.0000	0.0475	0.0278	0.0130	0.0296
Lazy.IBK	0.0046	0.0420	0.0601	0.0397	0.1009	0.0177	0.0711	0.0495	0.0302	0.0130	0.0322
Meta.logitboost	0.0976	0.0381	0.0489	0.0367	0.0968	0.0144	0.0829	0.0416	0.0242	0.0113	0.0257
Mean	0.0496	0.0306	0.0546	0.0345	0.0884	0.0161	0.0662	0.0455	0.0273	0.0121	0.0290
Std	0.0415	0.0122	0.0097	0.0043	0.0139	0.0029	0.0309	0.0072	0.0052	0.0018	0.0055
Overall mean	0.0413										
Overall std	0.0123										
Byes.NaiveBayes	0.0028	0.0500	0.0879	0.0520	0.1094	0.0259	0.0151	0.0732	0.0435	0.0203	0.0463
Treed.HoeffdingTree	0.0580	0.0476	0.0805	0.0459	0.1064	0.0237	0.0498	0.0673	0.0399	0.0186	0.0425
Meta.filteredclassifier	0.0396	0.0496	0.0830	0.0484	0.1080	0.0244	0.0472	0.0693	0.0411	0.0192	0.0438
Rules.part	0.0138	0.0568	0.0976	0.0499	0.1104	0.0288	0.0319	0.0811	0.0483	0.0226	0.0515
Tree.j48	0.0368	0.0546	0.0952	0.0494	0.1097	0.0280	0.0344	0.0792	0.0471	0.0220	0.0502
RandomTree	0.0000	0.0548	0.1029	0.0470	0.1090	0.0303	0.0263	0.0930	0.0435	0.0316	0.0463
Rule. Decision Table:	0.2890	0.0376	0.0513	0.0327	0.0935	0.0151	0.0804	0.0435	0.0254	0.0119	0.0270
Trees.LMT	0.2862	0.0393	0.0562	0.0357	0.0968	0.0165	0.0753	0.0475	0.0278	0.0130	0.0296
Mean	0.0908	0.0488	0.0818	0.0451	0.1054	0.0241	0.0450	0.0693	0.0396	0.0199	0.0422
Std	0.1231	0.0071	0.0189	0.0070	0.0065	0.0056	0.0231	0.0167	0.0085	0.0061	0.0091
Overall mean	0.0556										
Overall std	0.0211										

Therefore, the internal and external GDM rank is valid and undergoes systematic ranking.

Research limitation and future study

The proposed evaluation and benchmarking framework can address the evaluation and benchmarking issues for multiclass classification models. However, it cannot deal with classification models that work under multi-labelled or hierarchical cases because the evaluation criteria used for evaluation and benchmarking the multi-labelled or hierarchical cases are different and the procedures to calculate those criteria are different. The future study directions are as follows:

- The proposed framework can evaluate and benchmark the multiclass classification models that classify other types of leukaemia.
- The new framework can be applied for classification models with applications that involve the use of multi-labelled or hierarchical classification models through proposing new decision matrices that include related evaluation criteria for multi-labelled classification models or hierarchical classification models.

Conclusion

Studies related to the automated detection and classification of acute leukaemia have been notably increasing. Nevertheless, studies relevant to the evaluation and benchmarking of automated detection and classification tasks with unaddressed limitations are scarce. Several aspects are associated with the evaluation and benchmarking aimed for automated detection and classification. Such aspects warrant further analysis and investigation. Towards this end, comprehensive review and research on automated classification of acute leukaemia have been done while considering its evaluation and benchmarking aspects. The aim for the latter was to identify open challenges, research issues and gaps linked to the process of evaluation and benchmarking. After a thorough review of studies, a serious gap was identified. The gap resides in the failure of previous studies to perform a

process of evaluation and benchmarking for all major detection and classification requirements. Evaluation and benchmarking were partially performed, which render incomplete results because they failed to reflect the overall performance for detection and classification. Such weakness raises a challenge for comparing numerous systems or models for the detection and classification to determine which of the system or model is the best because the evaluation criteria vary and are incomplete. Moreover, all the major criteria and sub-criteria aimed for benchmarking multiclass detection and classification were reviewed. Towards addressing challenges, resolving issues and fulfilling the research gap, we proposed an evaluation and benchmarking framework based on MCDM techniques. Its goal is to evaluate and benchmark the acute leukaemia multiclass classification models. The description of the procedures and steps of the proposed framework are described. Construct decision matrix was based on crossover between evaluation criteria and 22 multiclass classification models. The proposed framework for evaluation and benchmarking are developed based on an integration of BWM and VIKOR. The ranking of classification models results are based on three experts' opinions on criterion preference. Firstly, the VIKOR was applied in the individual context to provide ranking for each expert, though the results show variances among the three experts' ranking. Therefore, VIKOR with GDM was applied, including internal and external aggregating methods. By contrast, internal and external aggregations have shown almost similar performance. Lastly, the validation for the results has been achieved objectively in this research. The statistical results indicate that the multiclass classification models ranking results based on internal and external aggregation GDM undergo a systematic ranking.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institution and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the study.

Appendix 1 pairwise comparisons

Section 1: Expert questionnaire

Dear Dr.,

The aim behind this questionnaire is to compare preferences between evaluation metrics of multiclass classification models of acute leukaemia for determining the importance for each metric. This questionnaire is a part of the research activities at Universiti Pendidikan Sultan Idris (UPSI)/Malaysia.

Background:

Name:

Years of experience:

E-Mail:

Position:

Prior to answering the questions, understanding the criteria assessed is important in arriving at a decision.

The criteria that usage for measurement the performance of a trained model on the test dataset. The evaluation criteria of acute leukaemia were divided into two main groups, namely, (1) reliability group, (2) time complexity;

The reliability group includes four subgroups of criteria, namely, (1) matrix of parameters has four metrics (i.e., confusion matrix: True positive, True negative, False negative, False positive), relationship of parameters has five metrics (i.e., Average Accuracy, Precision (Micro), Precision (Macro), Recall (Macro),, behaviour of parameters (F-score) and Error rate. The following Fig. 6 illustrates the levels:

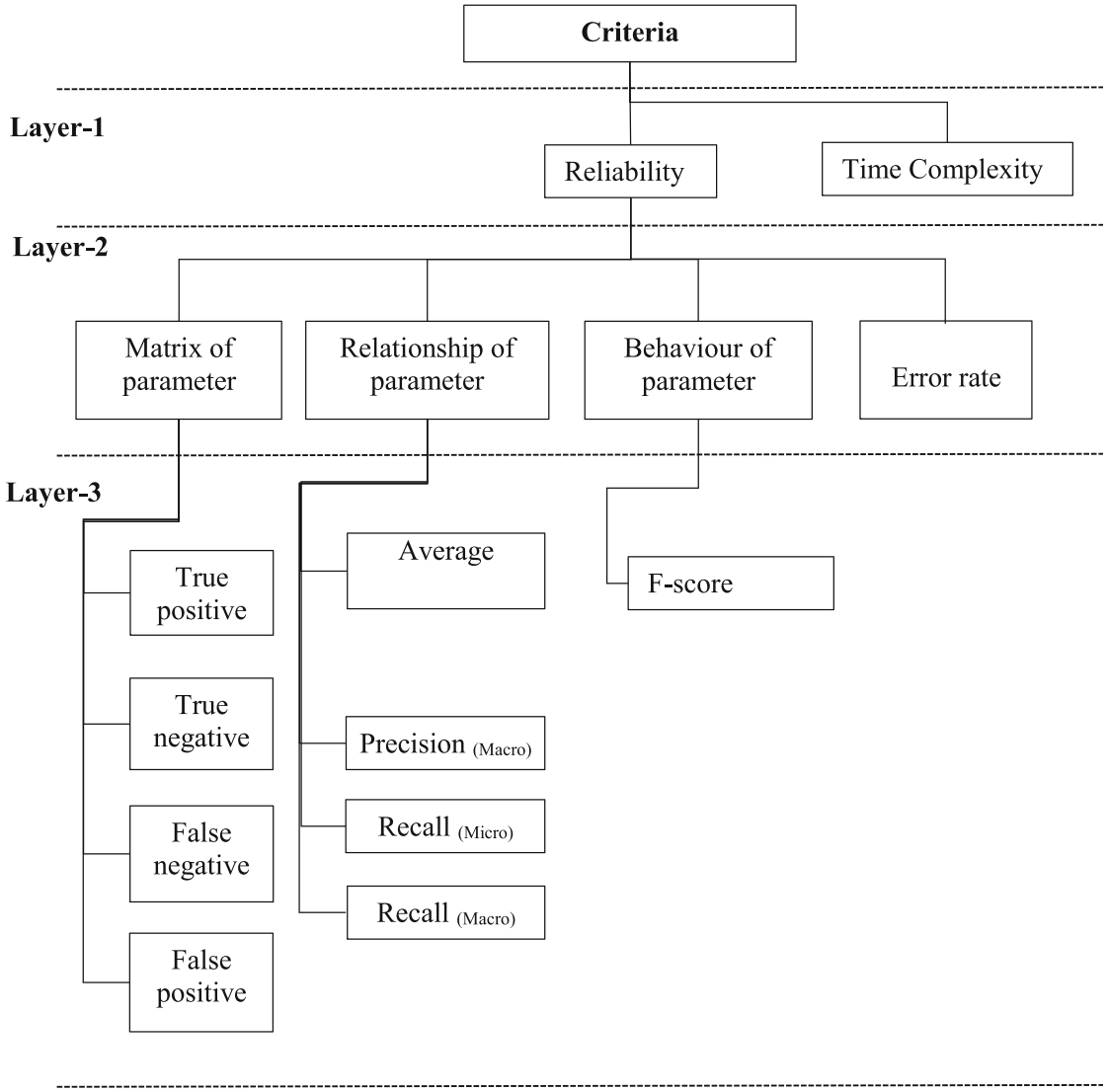

Fig. 6 illustrates the levels of evaluation criteria for multiclass classification models

Comparison questions

Comparison measurement scale

The comparisons (relative importance) of each criterion are measured according to a numerical scale from 1 to 9. These relative scales (1 to 9), as shown in Table 8, Please use this scale in comparison.

Table 8 Comparison measurement scale

Intensity of Importance	Definition
1	Equal importance
3	Moderately more important
5	Strongly more important
7	Very strongly more important
9	Extremely more important
2,4,6,8	Intermediate values

1. **Main Criteria**

A. **Reliability:** the degree of quality or state of being fit to be reliable value for any parameter. It is considered one of the main criteria in our study. This criterion includes four subsections will discuss in the next stage.

B. **Time Complexity:** is the time consumed by the input and output sample images, that's mean is the time required to complete the classification task of that algorithm.

Questions

1.1. Could you indicate, which of these two criteria you find is the MOST important and which one you find the LEAST important by marking the box? Please in Table 9, marking the cell of in front of the MOST important criterion and marking the cell of in front of the LEAST important criterion.

Table 9 Comparison to determine the most and least important criteria

Main Criteria	Most Important	Least Important
Reliability Group		
Time Complexity Group		

You have selected X criterion as the most important criterion.

1.2. Please determine your preference of this criterion (X) over the other least important criterion by using 1 to 9 measurement scale.

Please write the X criterion that you selected as most important criteria in green cell and the least important criterion in the grey cell in Table 10, and then write your preferences value.

Table 10 Comparison to determine the preference of most important criterion over other criteria

Criteria / Most Important	

2. **The sub-criteria (Level 2)**

A. **Matrix of parameter**:
It provides the statistics for the number of correct and incorrect predictions made by a classification system compared with the actual classifications of the samples in the test data

B. **Relationship of parameter:**
Relationship of parameters also included three parameters that are more important criteria typically used to measure the quality ratio for any case will discuss in the next
stage.

C. **Behaviour of parameter:**
Behaviour of parameters (f-score) that is to measure average harmonic mean and geometric for precision and recall perimeter will discuss in the next stage.

D. **Error rate**
Error rate within dataset: Basically, the procedure of dataset is to obtain the minimum error rate of the data during the implementation process of the training and validation applied in machine learning.

Questions

2.1. Could you indicate which one of these criteria (**sub-criteria (Level 2)**) consider the MOST important and which one you find the LEAST important? Please in Table 11, marking the cell of in front of the MOST important criterion and marking the cell of in front of the LEAST important criterion.

Table 11 Comparison to determine the most and least important criteria in level 2 of criteria

Sub-Criteria level 2	Most Important	Least Important
Matrix of parameter (Confusion matrix)		
Relationship of parameter		
Behaviour of parameter		
Error rate		

You have selected X criterion as the MOST important criterion and Y criterion as the LEAST important criterion

2.2. Please determine your preference of the criterion (X) over the other criteria by using 1 to 9 measurement scale.

Please write the X criterion that you selected as most important criterion in green cell and the other criteria in the grey cells in Table 12, and then write your preferences value.

2.3. You have selected Y criterion as the LEAST important criterion.

Please determine your preference of all criteria over the Y criteria that you selected as LEAST

Table 12 Comparison to determine the preference of most important criterion over the other criteria in level 2 of criteria

Most Important \ Criteria			

important criterion by using 1 to 9 measurement scale.

Please write the Y criterion that you selected as LEAST important criteria in green cell and the other criteria in the grey cells in Table 13, and then write your preferences value.

Table 13 Comparison to determine the preference of all criteria over the least important criterion in level 2 of criteria

Criteria \ Least Important	

3. **The sub-criteria (A) of Matrix of parameter (level 3)**

True positive	The number of elements correctly classified as positive by the test. When cancer cells are correctly identified
True negative	The number of elements correctly classified as negative by the test. When non-cancer cells are correctly identified
False positive	The number of elements classified as positive by the test, but they are not. When non-cancer cells are identified as cancerous
False negative	The number of elements classified as negative by the test, but they are not. When cancer cells are identified as noncancerous

Questions

3.1. Could you indicate which one of these criteria (**sub-criteria A(Level 3)**) consider the MOST important and which one you find the LEAST important? Please in Table 14, marking the cell of in front of the MOST important criterion and marking the cell of in front of the LEAST important criterion.

Table 14 Comparison to determine the most and least important criteria in the sub-criteria A level 3 of criteria

Sub-Criteria of matrix of parameter in level 3	Most Important	Least Important
True positive		
True negative		
False positive		
False negative		

Table 15 Comparison to determine the preference of most important criterion over the other criteria in the sub-criteria A level 3 of criteria

Most Important \ Criteria			

You have selected X criterion as the MOST important criterion and Y criterion as the LEAST important criterion

3.2. Please determine your preference of the criterion (X) over the other criteria by using 1 to 9 measurement scale.

Please write the X criterion that you selected as most important criterion in green cell and the other criteria in the grey cells in Table 15, and then write your preferences value.

3.3. You have selected Y criterion as the LEAST important criterion.

Please determine your preference of all criteria over the Y criteria that you selected as LEAST important criterion by using 1 to 9 measurement scale.

Please write the Y criterion that you selected as LEAST important criterion in green cell and the other criteria in the grey cells in Table 16, and then write your preferences value.

Table 16 Comparison to determine the preference of all criteria over the least important criterion in the sub-criteria A level 3 of criteria

Criteria \ Least Important	

4. **The sub-criteria (B) of Relationship of parameter in (level 3)**

Average Accuracy	The average effectiveness of all classes
$\text{Precision}_{(micro)}$	is used to measure the positive patterns that are correctly predicted from the total predicted patterns in a positive class (Agreement of the data class labels with those of a classifiers)
$\text{Precision}_{(macro)}$	Is an average per-class agreement of the data class labels with those of a classifier (An average per-class agreement of the data class with those of a classifiers).
$\text{Recall}_{(Macro)}$	Recall is used to measure the fraction of positive patterns that are correctly classified

Questions

4.1. Could you indicate which one of these criteria (**sub-criteria B (Level 3)**) consider the MOST important and which one you find the LEAST important? Please in Table 17, marking the cell of in front of the MOST important criterion and marking the cell of in front of the LEAST important criterion.

Table 17 Comparison to determine the most and least important criteria in the sub-criteria B level 3 of criteria

Sub-Criteria of Relationship of parameter in level 3	Most Important	Least Important
Average Accuracy		
Precision(micro)		
Precision(macro)		
Recall (Macro)		

X criterion selected as the best criterion and Y criterion as the LEAST important criterion

4.2. Determine your own preference of the criterion (X) compare the other criteria by using 1 to 9 measurement scale.

Please write the X criterion that you selected as most important criterion in green cell and the other criteria in the grey cells in Table 18, and then write your preferences value.

Table 18 Comparison to determine the preference of most important criterion over the other criteria in the sub-criteria B level 3 of criteria

Most Important \ Criteria				

4.3. Y criterion selected as the worst criterion.
Determine your own preference of all criteria compare with Y criterion that you selected as worst criterion by using 1 to 9 measurement scale.

Please write the Y criterion that you selected as LEAST important criterion in green cell and the other criteria in the grey cells in Table 19, and then write your preferences value.

Table 19 Comparison to determine the preference of all criteria over the least important criterion in the sub-criteria B level 3 of criteria

Criteria \ Least Important	

Should you have any inquiry or wish to know the result please contact:

Mohammed Assim Mohammed Ali
Email: Mohammed.asum@gmail.com
Mobile phone: 0060189810357

……. Thanks for Your Time …….

Section 2: List of experts

Table 20 List of experts involved in the pairwise questionnaire

Name	Work place / Country	Years of experience	Date
Dr. Dr. Marina Sokolova	University of Ottawa / Canada	More than 10 years	11/7/2018
Dr. Hishem Felouat	University Saad Dahlab Blida/ Algeria	More than 5 years	28/5/2018
Mazin A. Mohammed	University of Anbar / Iraq	More than 7 years	31/7/2018

Appendix 2 results of the BWM method for second and third experts

Table 21 The results of the BWM method for weight preferences of the criteria of evaluation and benchmarking the multiclass classification (second expert)

Expert 2							
Level 1 of Criteria							
("List of criteria")	("Best criterion")	("Other Criteria")	("Scores")	("Worst criterion")	("Other criteria")	("Scores")	Weight
R	R	TC	1				0.500
TC		–					0.500
Consistency: 0							
Level 2 of Criteria: sub criteria of Reliability							
List of criteria	Best criterion	Other Criteria	Scores	Other Criteria	Worst criterion	Scores	Weight
MOP	ER	ROP	3	ROP	BOP	3	0.147
ROP		BOP	7	ER		7	0.244
BOP		MOP	5	MOP		6	0.054
ER		–	–			–	0.555
Consistency: 0.048							
Level 3 of Criteria: sub criteria of Matrix of parameter							
List of criteria	Best criterion	Other Criteria	Scores	Other criteria	Worst criterion	Scores	Weight
TP	TP	TN	3	TN	FN	3	0.562
TN		FP	5	FP		5	0.237
FP		FN	7	TP		7	0.142
FN		–		–			0.059
Consistency: 0.040							
Level 3 of Criteria: sub criteria of Relationship of parameter							
List of criteria	Best criterion	Other Criteria	Scores	Worst criterion	Other criteria	Scores	Weight
Ave Accuracy	$Precision_{\mu}$	Ave Accuracy	5	$Precision_{\mu}$	Ave Accuracy	5	0.093
$Precision_{\mu}$		$Precision_M$	4	$Precision_M$		3	0.569
$Precision_M$		$Recall_M$	4	$Recall_M$		3	0.169
$Recall_M$							0.169
		–		–			
Consistency: 0.047							

Table 22 The results of the BWM method measurement for weight preferences of the evaluation and benchmarking for multiclass classification (Third expert)

Expert 3							
Level 1 of Criteria: Main Criteria							
List of criteria	Best criterion	Other Criteria	Scores				Weight
R	R	TC	4				0.800
TC		–					0.200
Consistency: 0							
Level 2 of Criteria: sub criteria of Reliability							
List of criteria	Best criterion	Other Criteria	Scores	Other criteria	Worst criterion	Scores	Weight
MOP	ROP	MOP	5	MOP	ER	2	0.149
ROP		BOP	3	BOP		5	0.537
BOP		ER	7	ROP		7	0.249
ER		–		–			0.065
Consistency: 0.065							
Level 3 of Criteria: sub criteria of Matrix of parameter							
List of criteria	Best criterion	Other Criteria	Scores	Other criteria	Worst criterion	Scores	Weight
TP	FN	TP	1	TP	TN	1	0.306
TN		TN	2	FN		2	0.130
FP		FP	1	FP		1	0.217
FN		–		–			0.348
Consistency: 0.087							
Level 3 of Criteria: sub criteria of Relationship of parameter							
List of criteria	Best criterion	Other Criteria	Scores	Worst criterion	Other criteria	Scores	Weight
Ave Accuracy	$Precision_{\mu}$	Ave Accuracy	3	$Precision_{\mu}$	Ave Accuracy	3	0.154
$Precision_{\mu}$		$Precision_M$	2	$Precision_M$		1	0.359
$Precision_M$		$Recall_M$	1	$Recall_M$		1	0.231
$Recall_M$		–	–	–		–	0.256
Consistency: 0.103							

Appendix 3 results of VIKOR for second and third experts

Table 23 Ranking results based on the second experts' weights

Machine learning	Q	Order
Rule.zero	0.0940	4
BayesNet	0.3616	20
Bayes.NaiveByesUpdateable	0.3069	16
Lazy.IBK	0.1614	8
Meta.AdaboostM1	0.3592	19
Meta.Bagging	0.3073	17
Meta.filteredclassifier	0.1516	7
Meta.logitboost	0.3371	18
Tree.j48	0.1116	5
REPTree	0.2045	12
RandomTree	0.0514	2
RandomForest	0.2772	15
Rule. Decision Table	1	22
Rules.part	0.0696	3
Meta.RandomCommittee	0.1812	10
Trees.LMT	0.9894	21
Treed.HoeffdingTree	0.1856	11
Kstar	0.1334	6
Functions.Smo	0.2511	14
Functions.SIMPLE.Logistic	0.1774	9
Byes.NaiveBayes	0	1
Decision Stump	0.2221	13

Table 24 Ranking results based on the third experts' weights

Machine learning	Q	Order
Rule.zero	0.5062	5
BayesNet	0.0170	2
Bayes.NaiveByesUpdateable	0	1
Lazy.IBK	0.8040	14
Meta.AdaboostM1	0.5092	6
Meta.Bagging	0.5836	9
Meta.filteredclassifier	0.9472	16
Meta.logitboost	0.7856	13
Tree.j48	0.9933	21
REPTree	0.5362	8
RandomTree	0.9848	20
RandomForest	0.4437	3
Rule. Decision Table	0.9591	18
Rules.part	0.9956	22
Meta.RandomCommittee	0.6587	10
Trees.LMT	0.9729	19
Treed.HoeffdingTree	0.9327	15
Kstar	0.5130	7
Functions.Smo	0.7596	11
Functions.SIMPLE.logistic	0.7799	12
Byes.NaiveBayes	0.9489	17
Decision Stump	0.4688	4

References

1. Salman, O., Zaidan, A., Zaidan, B., Naserkalid, and Hashim, M., Novel methodology for triage and prioritizing using "big data" patients with chronic heart diseases through telemedicine environmental.*Int. J. Inf. Technol. Decis. Mak.* 16(05):1211–1245, 2017.
2. Kalid, N. et al., Based on real time remote health monitoring systems: A new approach for prioritization "large scales data" patients with chronic heart diseases using body sensors and communication technology.*J. Med. Syst.* 42(4):69, 2018.
3. Mohsin, A. H. et al., Based medical systems for patient's authentication: Towards a new verification secure framework using CIA standard.*J. Med. Syst.* 43(7):192, 2019.
4. Mohsin, A. H. et al., Real-time medical systems based on human biometric steganography: A systematic review.*J. Med. Syst.* 42(12):245, 2018.
5. Mohsin, A. H. et al., Real-time remote health monitoring systems using body sensor information and finger vein biometric verification: A multi-layer systematic review.*J. Med. Syst.* 42(12):238, 2018.
6. Albahri, O. S. et al., Systematic review of real-time remote health monitoring system in triage and priority-based sensor technology: Taxonomy, open challenges, motivation and recommendations.*J. Med. Syst.* 42(5), 2018.
7. Abdulnabi, M. et al., A distributed framework for health information exchange using smartphone technologies.*J. Biomed. Inform.* 69:230–250, 2017.
8. Zaidan, A. A. et al., Challenges, alternatives, and paths to sustainability: Better public health promotion using social networking pages as key tools.*J. Med. Syst.* 39(2):7, 2015.
9. Mat Kiah, M. L. et al., Design and develop a video conferencing framework for real-time telemedicine applications using secure group-based communication architecture.*J. Med. Syst.* 38(10):133, 2014.
10. Shuwandy, M. L. et al., Sensor-based mHealth authentication for real-time remote healthcare monitoring system: A multilayer systematic review.*J. Med. Syst.* 43(2):33, 2019.
11. Talal, M. et al., Smart home-based IoT for real-time and secure remote health monitoring of triage and priority system using body sensors: Multi-driven systematic review.*J. Med. Syst.* 43(3):42, 2019.
12. Zaidan, B. B. et al., A security framework for Nationwide health information exchange based on telehealth strategy.*J. Med. Syst.* 39(5):51, 2015.
13. Hussain, M. et al., The landscape of research on smartphone medical apps: Coherent taxonomy, motivations, open challenges and recommendations.*Comput. Methods Prog. Biomed.* 122(3):393–408, 2015.
14. Zaidan, B. B. et al., Impact of data privacy and confidentiality on developing telemedicine applications: A review participates opinion and expert concerns.*Int. J. Pharmacol.* 7(3):382–387, 2011.
15. Kiah, M. L. M. et al., MIRASS: Medical informatics research activity support system using information mashup network. J. Med. Syst. 38(4):37, 2014.
16. Mohsin, A. H. et al., Based Blockchain-PSO-AES techniques in finger vein biometrics: A novel verification secure framework for patient authentication.*Comput. Stand. Interfaces*, 2019.
17. Hussain, M. et al., Conceptual framework for the security of mobile health applications on android platform.*Telematics Inform.* 35(5):1335, 2018.
18. Hussain, M. et al., A security framework for mHealth apps on android platform.*Comput. Secur.* 75:191–217, 2018.

19. Iqbal, S. et al., Real-time-based E-health systems: Design and implementation of a lightweight key management protocol for securing sensitive information of patients.*Health Technol. (Berl)*: 1–19, 2018.
20. Alanazi, H. O. et al., Meeting the security requirements of electronic medical records in the ERA of high-speed computing.*J. Med. Syst.* 39(1):165, 2015.
21. Nabi, M. S. A. et al., Suitability of using SOAP protocol to secure electronic medical record databases transmission.*Int. J. Pharmacol.* 6(6):959–964, 2010.
22. Kiah, M. L. M. et al., An enhanced security solution for electronic medical records based on AES hybrid technique with SOAP/XML and SHA-1.*J. Med. Syst.* 37(5):9971, 2013.
23. Nabi, M. S. et al., Suitability of adopting S/MIME and OpenPGP email messages protocol to secure electronic medical records. In: *Second International Conference on Future Generation Communication Technologies (FGCT 2013)*, 2013, 93–97.
24. Kiah, M. L. M. et al., Open source EMR software: Profiling, insights and hands-on analysis.*Comput. Methods Prog. Biomed.* 117(2):360–382, 2014.
25. Alsalem, M. A. et al., A review of the automated detection and classification of acute leukaemia: Coherent taxonomy, datasets, validation and performance measurements, motivation, open challenges and recommendations.*Comput. Methods Prog. Biomed.* 158:93–112, 2018.
26. Srisukkham, W., Zhang, L., Neoh, S. C., Todryk, S., and Lim, C. P., Intelligent leukaemia diagnosis with bare-bones PSO based feature optimization.*Appl. Soft Comput.* 56:405–419, 2017.
27. Labati, R. D., Piuri, V., Scotti, F., and Ieee, All-IDB: The acute lymphoblastic leukemia image database for image processing. In: *2011 18th Ieee International Conference on Image Processing*, 2011.
28. Lei, X., and Chen, Y., Multiclass classification of microarray data samples with flexible neural tree. In:*2012 Spring Congress on Engineering and Technology*, 2012, 1–4.
29. Agaian, S., Madhukar, M., and Chronopoulos, A. T., Automated screening system for acute myelogenous leukemia detection in blood microscopic images.*IEEE Syst. J.* 8:995–1004, 2014.
30. Mohapatra, S., Patra, D., and Satpathi, S., Image analysis of blood microscopic images for acute leukemia detection. In:*2010 International Conference on Industrial Electronics, Control and Robotics*, 2010, 215–219.
31. Bagasjvara, R. G., Candradewi, I., Hartati, S., and Harjoko, A., Automated detection and classification techniques of acute leukemia using image processing: A review. In:*2016 2nd International Conference on Science and Technology-Computer (ICST)*, 2016, 35–43.
32. Rawat, J., Singh, A., Bhadauria, H. S., and Virmani, J., Computer aided diagnostic system for detection of leukemia using microscopic images.*Procedia Computer Science* 70:748–756, 2015.
33. Snousy, M. B. A., El-Deeb, H. M., Badran, K., and Khlil, I. A. A., Suite of decision tree-based classification algorithms on cancer gene expression data.*Egyptian Informatics Journal* 12:73–82, 2011.
34. Goutam, D., and Sailaja, S., Classification of acute myelogenous leukemia in blood microscopic images using supervised classifier. In:*2015 IEEE International Conference on Engineering and Technology (ICETECH)*, 2015, 1–5.
35. Mishra, S., Majhi, B., Sa, P. K., and Sharma, L., Gray level co-occurrence matrix and random forest based acute lymphoblastic leukemia detection.*Biomedical Signal Processing and Control* 33: 272–280, 2017.
36. Nguyen, T., and Nahavandi, S., Modified AHP for gene selection and Cancer classification using Type-2 fuzzy logic.*IEEE Trans. Fuzzy Syst.* 24:273–287, 2016.
37. Hossin, M., and Sulaiman, M., A review on evaluation metrics for data classification evaluations.*International Journal of Data Mining & Knowledge Management Process* 5:1, 2015.
38. Sokolova, M., and Lapalme, G., A systematic analysis of performance measures for classification tasks.*Inf. Process. Manag.* 45: 427–437, 2009.
39. Krappe, S., Benz, M., Wittenberg, T., Haferlach, T., and Munzenmayer, C., Automated classification of bone marrow cells in microscopic images for diagnosis of leukemia: A comparison of two classification schemes with respect to the segmentation quality. In: Hadjiiski, L. M., Tourassi, G. D. (Eds),*Medical Imaging 2015: Computer-Aided Diagnosis*. Vol. 9414, 2015.
40. Cui, Y., Zheng, C.-H., Yang, J., and Sha, W., Sparse maximum margin discriminant analysis for feature extraction and gene selection on gene expression data.*Comput. Biol. Med.* 43:933–941, 2013.
41. Mohapatra, P., Chakravarty, S., and Dash, P. K., Microarray medical data classification using kernel ridge regression and modified cat swarm optimization based gene selection system.*Swarm and Evolutionary Computation* 28:144–160, 2016.
42. Wang, H.-Q., Wong, H.-S., Zhu, H., and Yip, T. T. C., A neural network-based biomarker association information extraction approach for cancer classification.*J. Biomed. Inform.* 42:654–666, 2009.
43. Zhang, L., and Xiaojuan, H., Multiple SVM-RFE for multi-class gene selection on DNA microarray data. In:*2015 International Joint Conference on Neural Networks (IJCNN)*, 2015, 1–6.
44. Yongqiang, D., Bin, H., Yun, S., Chengsheng, M., Jing, C., Xiaowei, Z. et al., Feature selection of high-dimensional biomedical data using improved SFLA for disease diagnosis. In:*2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2015, 458–463.
45. Salem, H., Attiya, G., and El-Fishawy, N., Gene expression profiles based human cancer diseases classification. In:*2015 11th International Computer Engineering Conference (ICENCO)*, 2015, 181–187.
46. Campos, L. M. d., Cano, A., Castellano, J. G., and Moral, S., Bayesian networks classifiers for gene-expression data. In:*2011 11th International Conference on Intelligent Systems Design and Applications*, 2011, 1200–1206.
47. Bhattacharjee, R., and Saini, L. M., Detection of acute lymphoblastic leukemia using watershed transformation technique. In: *2015 International Conference on Signal Processing, Computing and Control (ISPCC)*, 2015, 383–386.
48. Chandra, B., and Gupta, M., Robust approach for estimating probabilities in Naïve–Bayes classifier for gene expression data.*Expert Syst. Appl.* 38:1293–1298, 2011.
49. Singhal, V., and Singh, P., Local binary pattern for automatic detection of acute lymphoblastic leukemia. In:*2014 Twentieth National Conference on Communications (NCC)*, 2014, 1–5.
50. Rashid, S., and Maruf, G. M., An adaptive feature reduction algorithm for cancer classification using wavelet decomposition of serum proteomic and DNA microarray data. In:*2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, 2011, 305–312.
51. Ludwig, S. A., Jakobovic, D., and Picek, S., Analyzing gene expression data: Fuzzy decision tree algorithm applied to the classification of cancer data. In:*2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, 2015, 1–8.
52. Saritha, M., Prakash, B. B., Sukesh, K., and Shrinivas, B., Detection of blood cancer in microscopic images of human blood samples: A review. In:*2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, 2016, 596–600.
53. Tai, W. L., Hu, R. M., Hsiao, H. C. W., Chen, R. M., and Tsai, J. J. P., Blood cell image classification based on hierarchical SVM. In:

2011 IEEE International Symposium on Multimedia, 2011, 129–136.
54. Kumar, P. G., Aruldoss Albert Victoire, T., Renukadevi, P., and Devaraj, D., Design of fuzzy expert system for microarray data classification using a novel genetic swarm algorithm.*Expert Syst. Appl.* 39:1811–1821, 2012.
55. He, Y., and Hui, S. C., Exploring ant-based algorithms for gene expression data analysis.*Artif. Intell. Med.* 47:105–119, 2009.
56. Yusen, Z., and Liangyun, R., Two feature selections for analysis of microarray data. In:*2010 IEEE Fifth International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA)*, 2010, 1259–1262.
57. Rosa, J. L. D., Magpantay, A. E. A., Gonzaga, A. C., and Solano, G. A., Cluster center genes as candidate biomarkers for the classification of leukemia. In:*IISA 2014, the 5th International Conference on Information, Intelligence, Systems and Applications*, 2014, 124–129.
58. Lu, X., Peng, X., Liu, P., Deng, Y., Feng, B., and Liao, B., A novel feature selection method based on CFS in cancer recognition. In: *2012 IEEE 6th International Conference on Systems Biology (ISB)*, 2012, 226–231.
59. Kumar, M., and Kumar Rath, S., Classification of microarray using MapReduce based proximal support vector machine classifier.*Knowl.-Based Syst.* 89:584–602, 2015.
60. Dash, S., Hill-climber based fuzzy-rough feature extraction with an application to cancer classification. In:*13th International Conference on Hybrid Intelligent Systems (HIS 2013)*, 2013, 28–34.
61. Wahbeh, A. H., Al-Radaideh, Q. A., Al-Kabi, M. N., and Al-Shawakfa, E. M., A comparison study between data mining tools over some classification methods.*Int. J. Adv. Comput. Sci. Appl.* Special Issue on Artificial Intelligence:18–26, 2011.
62. Rangra, K., and Bansal, D. K. L., Comparative study of data mining tools.*International Journal of Advanced Research in Computer Science and Software Engineering* 4(6), 2014.
63. Yas, Q. M., Zaidan, A. A., Zaidan, B. B., Rahmatullah, B., and Karim, H. A., Comprehensive insights into evaluation and benchmarking of real-time skin detectors: Review, open issues & challenges, and recommended solutions. Measurement 114:243–260, 2018.
64. Wang, Z., and Palade, V., A comprehensive fuzzy-based framework for Cancer microarray data gene expression analysis. In: *2007 IEEE 7th International Symposium on BioInformatics and BioEngineering*, 2007, 1003–1010.
65. Nazlibilek, S., Karacor, D., Ercan, T., Sazli, M. H., Kalender, O., and Ege, Y., Automatic segmentation, counting, size determination and classification of white blood cells.*Measurement* 55:58–65, 2014.
66. Bhattacharjee, R., and Saini, L. M., Robust technique for the detection of acute lymphoblastic leukemia. In:*2015 IEEE Power, Communication and Information Technology Conference (PCITC)*, 2015, 657–662.
67. Torkaman, A., Charkari, N. M., Aghaeipour, M., and Hajati, E., A recommender system for detection of leukemia based on cooperative game. In:*2009 17th Mediterranean Conference on Control and Automation*, 2009, 1126–1130.
68. Escalante, H. J., Montes-y-Gómez, M., González, J. A., Gómez-Gil, P., Altamirano, L., Reyes, C. A. et al., Acute leukemia classification by ensemble particle swarm model selection.*Artif. Intell. Med.* 55:163–175, 2012.
69. Madhloom, H. T., Kareem, S. A., and Ariffin, H., A robust feature extraction and selection method for the recognition of lymphocytes versus acute lymphoblastic leukemia. In:*2012 International Conference on Advanced Computer Science Applications and Technologies (ACSAT)*, 2012, 330–335.
70. Cornet, E., Perol, J. P., and Troussard, X., Performance evaluation and relevance of the CellaVision (TM) DM96 system in routine analysis and in patients with malignant hematological diseases.*Int. J. Lab. Hematol.* 30:536–542, 2008.
71. Rota, P., Groeneveld-Krentz, S., and Reiter, M., On automated flow cytometric analysis for MRD estimation of acute lymphoblastic Leukaemia: A comparison among different approaches. In: *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2015, 438–441.
72. Keeney, R. L., and Raiffa, H.,*Decisions with Multiple Objectives: Preferences and Value Trade-Offs*. Cambridge: Cambridge university press, 1993.
73. Zaidan, A., Zaidan, B., Al-Haiqi, A., Kiah, M. L. M., Hussain, M., and Abdulnabi, M., Evaluation and selection of open-source EMR software packages based on integrated AHP and TOPSIS.*J. Biomed. Inform.* 53:390–404, 2015.
74. Khatari, M. et al., Multi-criteria evaluation and benchmarking for active queue management methods: Open issues, challenges and recommended pathway solutions.*Int. J. Inf. Technol. Decis. Mak.*: S0219622019300039, 2019.
75. Zaidan, A. A. et al., Multi-criteria analysis for OS-EMR software selection problem: A comparative study.*Decis. Support. Syst.* 78: 15–27, 2015.
76. Zaidan, B. B. et al., A new digital watermarking evaluation and benchmarking methodology using an external group of evaluators and multi-criteria analysis based on 'large-scale data.*Softw. Pract. Exp.* 47(10):1365–1392, 2017.
77. Yas, Q. M. et al., Towards on develop a framework for the evaluation and benchmarking of skin detectors based on artificial intelligent models using multi-criteria decision-making techniques.*Int. J. Pattern Recognit. Artif. Intell.* 31(03):1759002, 2017.
78. Belton, V., and Stewart, T.,*Multiple Criteria Decision Analysis: An Integrated Approach*. Boston: Kluwer Academic Publishers, 2002.
79. Zaidan, B., Zaidan, A., Abdul Karim, H., and Ahmad, N., A new approach based on multi-dimensional evaluation and benchmarking for data hiding techniques.*Int. J. Inf. Technol. Decis. Mak.*:1–42, 2017.
80. Zaidan, B., and Zaidan, A., Software and hardware FPGA-based digital watermarking and steganography approaches: Toward new methodology for evaluation and benchmarking using multi-criteria decision-making techniques.*Journal of Circuits, Systems and Computers* 26(07):1750116, 2017.
81. Abdullateef, B. N., Elias, N. F., Mohamed, H., Zaidan, A., and Zaidan, B., An evaluation and selection problems of OSS-LMS packages.*SpringerPlus* 5(1):248, 2016.
82. Qader, M. A. et al., A methodology for football players selection problem based on multi-measurements criteria analysis.*Measurement* 111:38–50, 2017.
83. Rahmatullah, B. et al., Multi-complex attributes analysis for optimum GPS baseband receiver tracking channels selection. In:*2017 4th International Conference on Control, Decision and Information Technologies, CoDIT 2017*. Vol. 2017, 2017, 1084–1088.
84. Jumaah, F. M. et al., Technique for order performance by similarity to ideal solution for solving complex situations in multi-criteria optimization of the tracking channels of GPS baseband telecommunication receivers.*Telecommun. Syst.*:1–19, 2018.
85. Petrovic-Lazarevic, S., & Abraham, A., Hybrid fuzzy-linear programming approach for multi criteria decision making problems.*Neural Parallel & Scientific Comp.*, 11:53-68, 2003.
86. Malczewski, J.,*GIS and Multicriteria Decision Analysis*. New York: Wiley, 1999.
87. Alsalem, M., Zaidan, A., Zaidan, B., Hashim, M., Albahri, O., Albahri, A. et al., Systematic review of an automated multiclass

detection and classification system for acute Leukaemia in terms of evaluation and benchmarking, open challenges, issues and methodological aspects.*J. Med. Syst.* 42(11):204, 2018.

88. Yas, Q. M. et al., Comprehensive insights into evaluation and benchmarking of real-time skin detectors: Review, open issues & challenges, and recommended solutions.*Measurement* 114:243–260, 2018.
89. Zaidan, B. B., and Zaidan, A. A., Comparative study on the evaluation and benchmarking information hiding approaches based multi-measurement analysis using TOPSIS method with different normalisation, separation and context techniques.*Measurement* 117:277–294, 2018.
90. Zaidan, A. A. et al., A review on smartphone skin cancer diagnosis apps in evaluation and benchmarking: Coherent taxonomy, open issues and recommendation pathway solution.*Health Technol. (Berl).* 8(4):223–238, 2018.
91. Zionts, S., MCDM-if not a Roman numeral, then what?*Interfaces* 9:94–101, 1979.
92. Baltussen, R., and Niessen, L., Priority setting of health interventions: The need for multi-criteria decision analysis.*Cost effectiveness and resource allocation* 4:1, 2006.
93. Thokala, P., Devlin, N., Marsh, K., Baltussen, R., Boysen, M., Kalo, Z. et al., Multiple criteria decision analysis for health care decision making—An introduction: Report 1 of the ISPOR MCDA emerging good practices task force.*Value Health* 19:1–13, 2016.
94. Oliveira, M., Fontes, D. B., and Pereira, T., Multicriteria decision making: A case study in the automobile industry.*Annals of Management Science* 3:109, 2014.
95. Tariq, I. et al., MOGSABAT: A metaheuristic hybrid algorithm for solving multi-objective optimisation problems.*Neural Comput. & Applic.* 30:1–15, 2018.
96. Enaizan, O. et al., Electronic medical record systems: Decision support examination framework for individual, security and privacy concerns using multi-perspective analysis.*Health Technol.*, 1-18, 2018.
97. Salih, M. M. et al., Survey on fuzzy TOPSIS state-of-the-art between 2007–2017.*Comput. Oper. Res.*, 104:207–227, 2019.
98. Kalid, N. et al., Based real time remote health monitoring systems: A review on patients prioritization and related" big data" using body sensors information and communication technology.*J. Med. Syst.* 42(2):30, 2018.
99. Jumaah, F. M. et al., Decision-making solution based multi-measurement design parameter for optimization of GPS receiver tracking channels in static and dynamic real-time positioning multipath environment.*Measurement* 118:83–95, 2018.
100. Jadhav, A., and Sonar, R., Analytic hierarchy process (AHP), weighted scoring method (WSM), and hybrid knowledge based system (HKBS) for software selection: A comparative study. In: *2009 Second International Conference on Emerging Trends in Engineering & Technology*, 2009, 991–997.
101. Albahri, A. S. et al., Real-time fault-tolerant mHealth system: Comprehensive review of healthcare services, opens issues, challenges and methodological aspects.*J. Med. Syst.* 42(8):137, 2018 Springer US.
102. Albahri, O. S. et al., Real-time remote health-monitoring systems in a Medical Centre: A review of the provision of healthcare services-based body sensor information, open challenges and methodological aspects.*J. Med. Syst.* 42(9):164, 2018.
103. Talal, M. et al., Comprehensive review and analysis of anti-malware apps for smartphones.*Telecommun. Syst.*, 1-53, 2019.
104. Zaidan, A. A. et al., Based multi-agent learning neural network and Bayesian for real-time IoT skin detectors: A new evaluation and benchmarking methodology.*Neural Comput. & Applic.*, 2019.
105. Albahri, A. S. et al., Based multiple heterogeneous wearable sensors: A smart real-time health monitoring structured for hospitals distributor.*IEEE Access* 7:37269–37323, 2019.
106. Albahri, O. S. et al., Fault-tolerant mHealth framework in the context of IoT-based real-time wearable health data sensors.*IEEE Access* 7:50052–50080, 2019.
107. Whaiduzzaman, M., Gani, A., Anuar, N. B., Shiraz, M., Haque, M. N., and Haque, I. T., Cloud service selection using multicriteria decision analysis.*Sci. World J.* 2014:459375, 2014.
108. Aruldoss, M., Lakshmi, T. M., and Venkatesan, V. P., A survey on multi criteria decision making methods and its applications.*American Journal of Information Systems* 1:31–43, 2013.
109. Singh, A., & Malik, SK., Major MCDM techniques and their application-a review.*IOSR Journal of Engineering*, 4(5):15-25, 2014.
110. Opricovic, S., and Tzeng, G.-H., Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS.*Eur. J. Oper. Res.* 156:445–455, 2004.
111. Guo, S., and Zhao, H., Fuzzy best-worst multi-criteria decision-making method and its applications.*Knowl.-Based Syst.* 121:23–31, 2017.
112. Rezaei, J., Best-worst multi-criteria decision-making method.*Omega* 53:49–57, 2015.
113. Tavana, M., and Hatami-Marbini, A., A group AHP-TOPSIS framework for human spaceflight mission planning at NASA.*Expert Syst. Appl.* 38:13588–13603, 2011.
114. Zaidan, A. A., Zaidan, B. B., Albahri, O. S., Alsalem, M. A., Albahri, A. S., Yas, Q. M. et al., A review on smartphone skin cancer diagnosis apps in evaluation and benchmarking: Coherent taxonomy, open issues and recommendation pathway solution.*Heal. Technol.* 8:223–238, 2018.
115. Azeez, D., Ali, M. A. M., Gan, K. B., and Saiboon, I., Comparison of adaptive neuro-fuzzy inference system and artificial neutral networks model to categorize patients in the emergency department.*SpringerPlus* 2:416, 2013.
116. Ashour, O. M., and Okudan, G. E., Fuzzy AHP and utility theory based patient sorting in emergency departments. International Journal of Collaborative Enterprise 1:332–358, 2010.
117. Mills, A. F., A simple yet effective decision support policy for mass-casualty triage.*Eur. J. Oper. Res.* 253:734–745, 2016.
118. Adunlin, G., Diaby, V., and Xiao, H., Application of multicriteria decision analysis in health care: A systematic review and bibliometric analysis.*Health Expect.* 18:1894–1905, 2015.
119. Jumaah, F., Zadain, A., Zaidan, B., Hamzah, A., and Bahbibi, R., Decision-making solution based multi-measurement design parameter for optimization of GPS receiver tracking channels in static and dynamic real-time positioning multipath environment.*Measurement*, 118:83-95, 2018.
120. Yas, Q. M., Zaidan, A., Zaidan, B., Rahmatullah, B., and Karim, H. A., Comprehensive insights into evaluation and benchmarking of real-time skin detectors: Review, open issues & challenges, and recommended solutions.*Measurement*, 114:243-260, 2018.
121. Nilsson, H., Nordström, E.-M., and Öhman, K., Decision support for participatory forest planning using AHP and TOPSIS.*Forests* 7:100, 2016.
122. Kornyshova, E., and Salinesi, C., MCDM techniques selection approaches: State of the art. In:*2007 IEEE Symposium on Computational Intelligence in Multi-Criteria Decision-Making*, 2007, 22–29.
123. Kaya, İ., Çolak, M., and Terzi, F., Use of MCDM techniques for energy policy and decision-making problems: A review.*Int. J. Energy Res.* 42:2344–2372, 2018.
124. Wan Ahmad, W. N. K., Rezaei, J., Sadaghiani, S., and Tavasszy, L. A., Evaluation of the external forces affecting the sustainability of

oil and gas supply chain using best worst method.*J. Clean. Prod.* 153:242–252, 2017.
125. Gupta, H., and Barua, M. K., Supplier selection among SMEs on the basis of their green innovation ability using BWM and fuzzy TOPSIS.*J. Clean. Prod.* 152:242–258, 2017.
126. Rezaei, J., Best-worst multi-criteria decision-making method: Some properties and a linear model.*Omega* 64:126–130, 2016.
127. Yang, Q., Zhang, Z., You, X., and Chen, T., Evaluation and classification of overseas talents in China based on the BWM for intuitionistic relations.*Symmetry* 8:137, 2016.
128. Opricovic, S., and Tzeng, G.-H., Extended VIKOR method in comparison with outranking methods.*Eur. J. Oper. Res.* 178: 514–529, 2007.
129. Mahjouri, M., Ishak, M. B., Torabian, A., Abd Manaf, L., Halimoon, N., and Ghoddusi, J., Optimal selection of Iron and steel wastewater treatment technology using integrated multi-criteria decision-making techniques and fuzzy logic.*Process Saf. Environ. Prot.* 107:54–68, 2017.
130. Ren, J., Selection of sustainable prime mover for combined cooling, heat, and power technologies under uncertainties: An interval multicriteria decision-making approach.*Int. J. Energy Res.*, 42(8):2655-2669, 2018.
131. Gupta, H., Evaluating service quality of airline industry using hybrid best worst method and VIKOR.*J. Air Transp. Manag.* 68:35–47, 2018.
132. Serrai, W., Abdelli, A., Mokdad, L., and Hammal, Y., An efficient approach for web service selection. In:*2016 IEEE Symposium on Computers and Communication (ISCC)*, 2016, 167–172.
133. Shojaei, P., Seyed Haeri, S. A., and Mohammadi, S., Airports evaluation and ranking model using Taguchi loss function, best-worst method and VIKOR technique.*J. Air Transp. Manag.* 68:4–13, 2018.
134. Serrai, W., Abdelli, A., Mokdad, L., and Hammal, Y., Towards an efficient and a more accurate web service selection using MCDM methods.*J. Comput. Sci.* 22:253–267, 2017.
135. Pamučar, D., Petrović, I., and Ćirović, G., Modification of the best–worst and MABAC methods: A novel approach based on interval-valued fuzzy-rough numbers.*Expert Syst. Appl.* 91:89–106, 2018.
136. Tian, Z.-p., Wang, J.-q., and Zhang, H.-y., An integrated approach for failure mode and effects analysis based on fuzzy best-worst, relative entropy, and VIKOR methods.*Appl. Soft Comput.*, 72: 636-646, 2018.
137. Chiu, W.-Y., Tzeng, G.-H., and Li, H.-L., A new hybrid MCDM model combining DANP with VIKOR to improve e-store business.*Knowl.-Based Syst.* 37:48–61, 2013.
138. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P. et al., Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring.*Science* 286:531–537, 1999.
139. Zhou, C., Wan, L., and Liang, Y., A hybrid algorithm of minimum spanning tree and nearest neighbor for classifying human cancers. In:*Advanced Computer Theory and Engineering (ICACTE), 2010 3rd International Conference on*, 2010, V5-585–V5-589.
140. Chakraborty, S., Simultaneous cancer classification and gene selection with Bayesian nearest neighbor method: An integrated approach.*Computational Statistics & Data Analysis* 53:1462–1474, 2009.
141. Chunbao, Z., Liming, W., and Yanchun, L., A hybrid algorithm of minimum spanning tree and nearest neighbor for classifying human cancers. In:*2010 3rd International Conference on Advanced Computer Theory and Engineering(ICACTE)*, 2010, V5-585–V5-589.
142. Horng, J.-T., Wu, L.-C., Liu, B.-J., Kuo, J.-L., Kuo, W.-H., and Zhang, J.-J., An expert system to classify microarray gene expression data using gene selection by decision tree.*Expert Syst. Appl.* 36:9072–9081, 2009.
143. Garro, B. A., Rodríguez, K., and Vazquez, R. A., Designing artificial neural networks using differential evolution for classifying DNA microarrays. In:*2017 IEEE Congress on Evolutionary Computation (CEC)*, 2017, 2767–2774.
144. Al-Sahaf, H., Song, A., and Zhang, M., Hybridisation of genetic programming and nearest neighbour for classification. In:*2013 IEEE Congress on Evolutionary Computation*, 2013, 2650–2657.
145. Deegalla, S., and Boström, H., Improving fusion of dimensionality reduction methods for nearest neighbor classification. In:*2009 International Conference on Machine Learning and Applications*, 2009, 771–775.
146. Hasan, A., and Akhtaruzzaman, A. M., High dimensional microarray data classification using correlation based feature selection. In:*2012 International Conference on Biomedical Engineering (ICoBE)*, 2012, 319–321.
147. Huang, P. H., and Moh, T.-t., A non-linear non-weight method for multi-criteria decision making.*Ann. Oper. Res.* 248:239–251, 2017.
148. Aboutorab, H., Saberi, M., Asadabadi, M. R., Hussain, O., and Chang, E., ZBWM: The Z-number extension of best worst method and its application for supplier development.*Expert Syst. Appl.* 107:115–125, 2018.
149. Almahdi, E. M. et al., Based mobile patient monitoring systems: A prioritization framework using multi-criteria decision making techniques. *J. Med. Syst*. 43, 2019.
150. Almahdi, E. M. et al., Mobile patient monitoring systems from a benchmarking aspect: Challenges, open issues and recommended solutions. *J. Med. Syst*. 43, 2019.