Journal of Medical Systems (2018) 42: 204
https://doi.org/10.1007/s10916-018-1064-9

SYSTEMS-LEVEL QUALITY IMPROVEMENT

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

M. A. Alsalem [1] · A. A. Zaidan [1] · B. B. Zaidan [1] · M. Hashim [1] · O. S. Albahri [1] · A. S. Albahri [1] · Ali Hadi [1] · K. I. Mohammed [1]

Received: 18 July 2018 / Accepted: 6 September 2018 / Published online: 19 September 2018

Abstract

This study aims to systematically review prior research on the evaluation and benchmarking of automated acute leukaemia classification tasks. The review depends on three reliable search engines: ScienceDirect, Web of Science and IEEE Xplore. A research taxonomy developed for the review considers a wide perspective for automated detection and classification of acute leukaemia research and reflects the usage trends in the evaluation criteria in this field. The developed taxonomy consists of three main research directions in this domain. The taxonomy involves two phases. The first phase includes all three research directions. The second one demonstrates all the criteria used for evaluating acute leukaemia classification. The final set of studies includes 83 investigations, most of which focused on enhancing the accuracy and performance of detection and classification through proposed methods or systems. Few efforts were made to undertake the evaluation issues. According to the final set of articles, three groups of articles represented the main research directions in this domain: 56 articles highlighted the proposed methods, 22 articles involved proposals for system development and 5 papers centred on evaluation and comparison. The other taxonomy side included 16 main and sub-evaluation and benchmarking criteria. This review highlights three serious issues in the evaluation and benchmarking of multiclass classification of acute leukaemia, namely, conflicting criteria, evaluation criteria and criteria importance. It also determines the weakness of benchmarking tools. To solve these issues, multicriteria decision-making (MCDM) analysis techniques were proposed as effective recommended solutions in the methodological aspect. This methodological aspect involves a proposed decision support system based on MCDM for evaluation and benchmarking to select suitable multiclass classification models for acute leukaemia. The said support system is examined and has three sequential phases. Phase One presents the identification procedure and process for establishing a decision matrix based on a crossover of evaluation criteria and acute leukaemia multiclass classification models. Phase Two describes the decision matrix development for the selection of acute leukaemia classification models based on the integrated Best and worst method (BWM) and VIKOR. Phase Three entails the validation of the proposed system.

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

M. Hashim
mashitoh.hashim@fskik.upsi.edu.my

O. S. Albahri
osamahsh89@gmail.com

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

Ali Hadi
ali_hadi182@yahoo.com

K. I. Mohammed
khalid_ib81@yahoo.com

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

Keywords Acute leukaemia · Multiclass classification · Detection · Evaluation · Benchmarking · Evaluation criteria · MCDM

Introduction

The application of automation systems is very important in complex medical cases [1]. Automated detection and classification of acute leukaemia is necessary to provide patients with suitable treatment and mitigate its dangers. The rapid and accurate diagnosis of this type of cancer plays a core role in patient treatment and recovery [2–5]. Most automation systems for complex medical problems depend on machine learning techniques, where machine learning is one of the common scientific fields based on artificial intelligence concepts. Machine learning techniques can handle many issues related to acute leukaemia, such as diagnosis, detection and classification [6–8]. Hence, numerous studies [1, 9–12] confirmed the necessity of adopting automated systems and methods to deal with various issues related to acute leukaemia. They stated that these systems provide precision results with fast response. Accordingly, many investigations focused on proposing or enhancing the detection and classification methods for acute leukaemia. Others studies [11, 13–18] developed automated systems to manage issues related to acute leukaemia, and all these efforts attempted to provide optimal results regarding acute leukaemia classification and detection [16, 19]. Moreover, the automated classification of acute leukaemia became common in many hospitals and cancer specialist centres to overcome the limitations of manual analysis [16, 18, 19]. Despite all the benefits obtained from these systems, however, users began facing challenges in choosing an automated system that provides highly accurate results with the highest performance among many available alternatives [20]. The vast diversity among available classification systems for acute leukaemia makes it difficult for health organisations to decide on which system to use. Therefore, the administrations of health organisations encounter difficulty in evaluating and comparing automated classification systems for acute leukaemia to select the best system, especially as no single system is superior to the rest [15, 16, 20] and many suffer from a lack of accuracy and computational efficiency [21]. Conversely, the difficulty of evaluation and comparison arises due to the multiple criteria of evaluation and the conflict among them [22]. The evaluation and benchmarking of automated classification systems for serious medical cases such as acute leukaemia are crucial in the quest for ascertaining the optimal system [10]. Such a process is critical because the wrong classification system can cost health organisations loss of patient life, legal accountability and even financial costs if the system fails to live up to expectations. For instance, if the system incorrectly identifies non-cancer cells as cancerous, that outcome may have adverse effects on the patient's mental state, and he/she may need further surgery and diagnosis to determine whether he/she is cancer-free. The most serious case is when the system incorrectly identifies cancer cells as non-cancerous. Such an error is more important in this case because the existence of the disease will go unnoticed, appropriate therapy will not be implemented and then loss of life may transpire. Both cases will have a negative impact on the reputation and performance of healthcare organisations [189, 190, 194, 195]. Thus, identifying the most efficient technique to help health organisations in making right decisions on classification system selection is necessary [197, 201, 202, 209, 211]. Evaluating and benchmarking processes are required for selecting the best automated classification system among many available alternatives, especially since these systems are not cheap and related to human medical concerns [21]. A comprehensive review of literature is essential to highlight the automated multiclass classification systems for acute leukaemia, the benefits and characteristic of these systems from a wide bibliography and the challenges in the selection of the best classification systems resulting from the difficulty of evaluating and benchmarking these systems. The challenges and open issues in the selection of acute leukaemia classification systems need further study and analysis. The methodological aspects of providing a decision support system for the evaluation and benchmarking must also be emphasised to ensure the continuous provision of a better multiclass classification model for acute leukaemia through choosing the optimal classification system and overcoming related challenges. Figure 1 presents the framework of the literature review for this study. The remainder of this study consists of three parts. Part 2 provides a review and an in-depth and comprehensive analysis of past studies. Part 3 presents a discussion of the methodological aspects for our proposed decision support system. Part 4 provides the study conclusion.

Comprehensive review

A literature review is detailed in the following sections. All the steps and procedures in the protocol of the systematic review for evaluating and benchmarking the classification of acute leukaemia are described.

Systematic review protocol

This section presents the protocol of systematic review used in this study. The method of systematic review, information sources, selection of studies, search process, inclusion and exclusion criteria, data collection and literature taxonomy are described below.

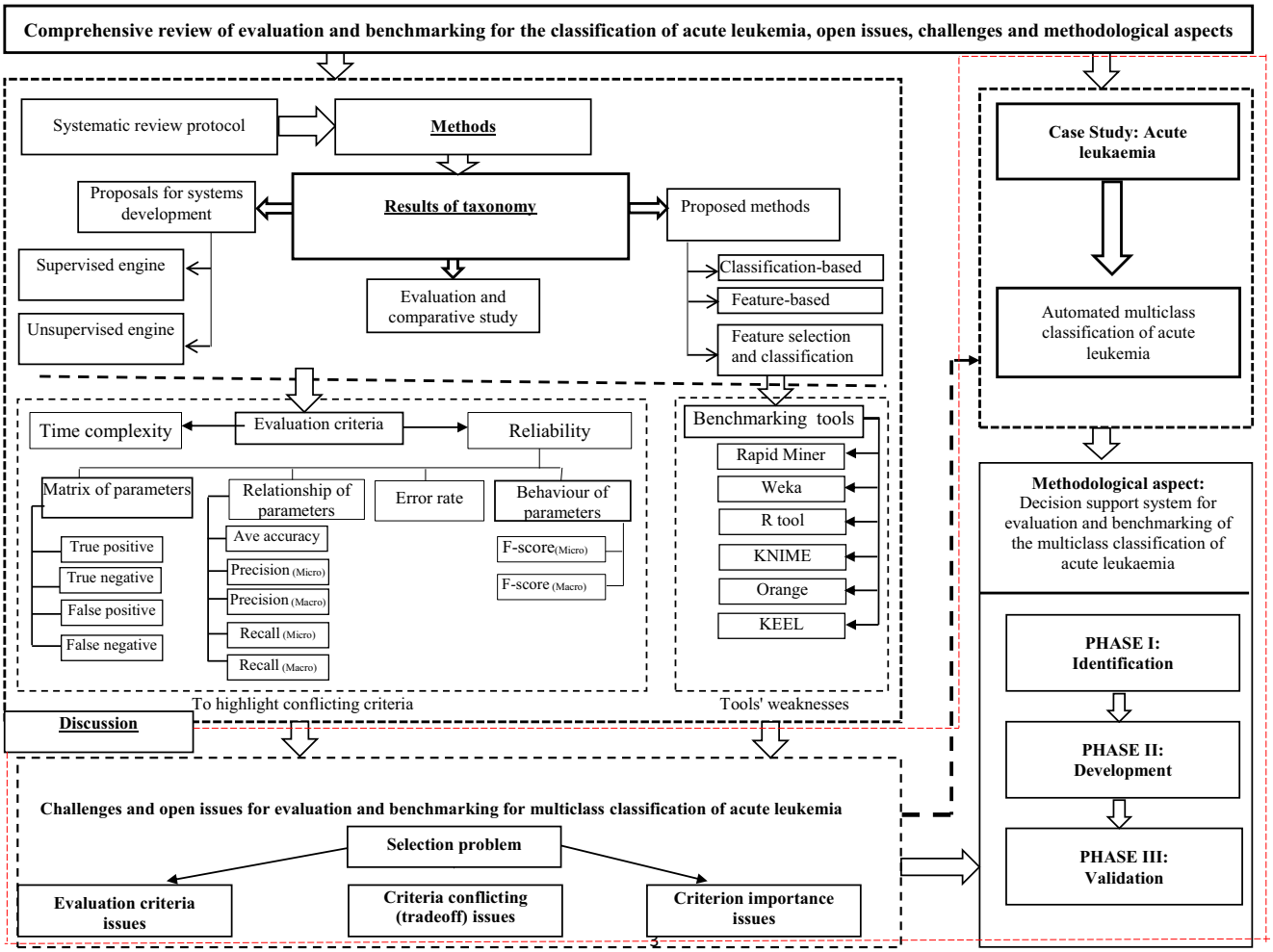

Fig. 1 Framework of literature review

Methods

This section describes the procedures in the search, collection, filtering and reading of articles. Three reliable indexes, namely, IEEE Xplore, ScienceDirect (SD) and Web of Science (WoS) were adopted in searching for the articles. These three indexes cover a wide range of journals and conference articles related to our study field. They are also characterised by ease of use and the capability to build simple and complex search queries. The research query was used on IEEE Xplore, SD and WoS. The main keywords formed the search query ('leukemia' OR 'leukemias' OR 'Leukaemia') along with the main terms of evaluation ('Evaluation' OR 'Assessment' OR 'Benchmark' OR 'measurement') with computerised and AI terms to limit the scope of search to only the articles that adopted automation methods ('AI' OR 'artificial intelligent' OR 'automated' OR 'Automation' OR 'Computerized'). The queries were run via the expert search form on these three databases. The search range covered only the articles and conference papers in the last 10 years. The query string is presented at the top of Fig. 2. Articles duplicated in the three selected databases were removed after the search process. Two iterations of intensive search were conducted in the selection of relevant articles. The first iteration focused on excluding irrelevant articles through scanning the titles and abstracts. The second one entailed intensive full-text reading for all the relevant articles. These two rounds were conducted using similar eligibility criteria. The third iteration emphasised screening the last set of studies to determine the evaluation criteria applied in the evaluation process of acute leukaemia classification tasks and excluded any study that did not utilise any of the identified evaluation criteria. The final set of articles was related to all studies which used any of the evaluation criteria for the binary and multiclass classification tasks of acute leukaemia. Note that review and survey studies that mentioned the evaluation criteria but did not actually use them were excluded in the third iteration. Important information used in writing this review which was extracted from the relevant articles during the full reading was saved into an Excel file. Only studies that fulfil the inclusion criteria listed in Fig. 2 were included. The exclusion conditions applied are as follows: non-English papers, non-English articles, studies which did

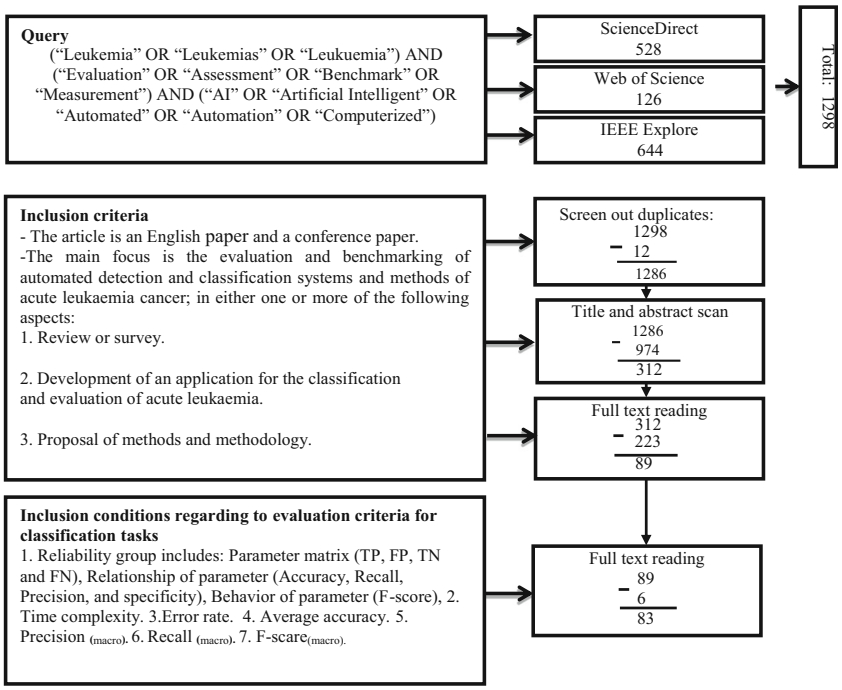

Fig. 2 Selection of studies, search query and inclusion criteria

not focus on acute leukaemia detection or classification, studies which focused only on segmentation of acute leukaemia images and the manual classification and analysis of acute leukaemia and articles that did not use any of the evaluation criteria. The list of relevant papers was organised in an Excel sheet file and EndNote library [198–200, 203, 204, 208]. Through full-text reading, the authors extracted numerous important highlights and information from the surveyed papers [183, 184, 188, 196]. The researchers were keen to extract important information and details that enabled them to reflect a detailed picture of all the features and aspects of acute leukaemia detection and classification with related evaluation and benchmarking processes. According to the purpose of this study, all relevant articles were classified into three categories and formulised into literature taxonomy. All the important highlights and information were placed on the body texts.

Results of taxonomy This part describes all the results from the initial search until the final results were reached. The first query searches in the three databases yielded 1298 papers, the largest proportion (644 papers: 49%) were from IEEE Xplore, followed by SD (528 papers: 41%), and then WoS (126 papers: 10%). The search period covered publications in the last 10 years. In the first scan, 12 duplicate articles were excluded and 974 were excluded after reading their title and abstract, which indicated that they did not meet the inclusion criteria. A total of 312 articles remained after the exclusion of irrelevant and duplicated articles. Using the eligibility criteria, 312 irrelevant papers were excluded via full-text reading, resulting in the semi-final set of 89 papers gathered through the inclusion criteria. The final set of 83 papers involved studies that used at least one of the identified evaluation criteria. The final set was analysed and used to develop the research taxonomy, which considers a wide perspective for automated detection and classification of acute leukaemia research and reflects the trends of usage of the evaluation criteria in this field.

With the final set of articles, three groups were organised that represented the main research directions in this domain. The taxonomy was achieved in two phases, the first one (Fig. 3, left section) included three research directions: the first direction included 56 articles that focused on proposed methods, the second one covered 22 articles that presented proposals for system development and the third direction included 5 papers on evaluation and comparison. The second phase (Fig. 2, right section) demonstrated all the criteria used for evaluating acute leukaemia classification.

Proposed methods The first category covers the research efforts on classification and feature selection/reduction. This category includes 56 articles that focused on enhancing the performance of classification or proposing new classification methods to deal with the binary or multiclass classification problem and improve feature selection/reduction. This category includes three groups of papers. The largest one has 33 articles that propose new classification methods or the enhancement of current approaches. The second one has 15 papers that deal with feature selection/reduction methods, and the last one has 8 papers that contribute to the classification tasks and the resolution of feature problems. The following section will describe these three groups of studies while emphasising the evaluation criteria used in each study.

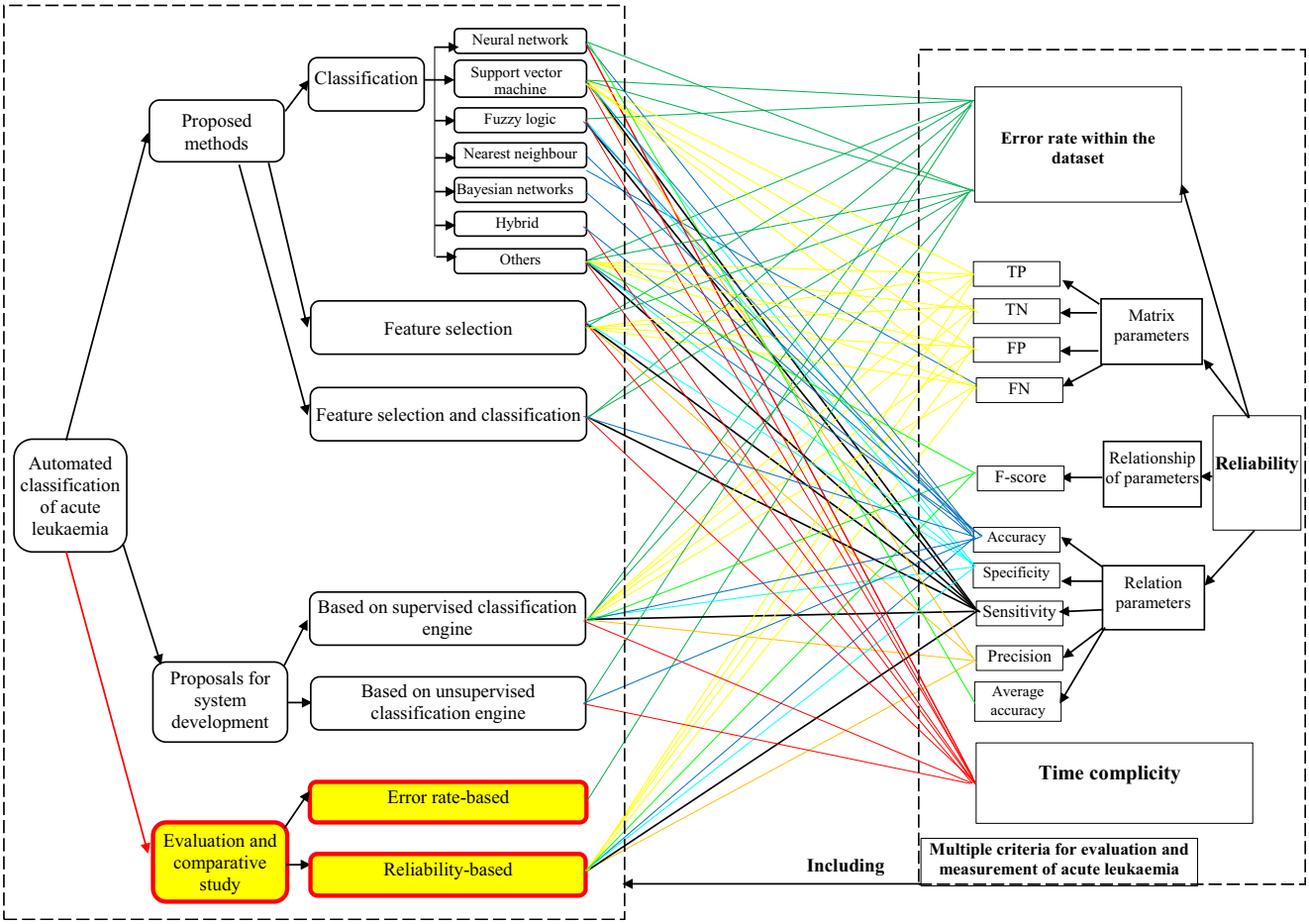

Fig. 3 Literature taxonomy and evaluation criteria

Firstly, this study depends on average accuracy and time for measuring the performance of the flexible neural multiclass algorithm in [23]. In [24], accuracy and training were used as performance indicators for the general regression of the neural network. Moreover, the neural network depends on the association data extraction method in [25] and relies on accuracy.

Secondly, accuracy and training were used in [26] to evaluate the proposed twin-space SVM. Accuracy–time criteria were used to boost an evolutionary support vector machine [27]. This study [28] was comprehensive in using evaluation criteria, accuracy, precision, specificity, sensitivity, true positive (tp), true negative (tn), false positive (fp), false negative (fn), and time complexity as validation measures and proposed a map-reduced-based proximal SVM method.

Thirdly, FUZZY-ARTMAP neural networks used in [29] depended on two evaluation criteria, accuracy and validation, to classify the WBCs. These evaluation criteria involving accuracy, specificity, sensitivity and time were utilised to evaluate the classification tasks in [22]. Another work [30] employed accuracy and time as performance metrics of fuzzy nearest neighbours, while [31] used accuracy-specificity and sensitivity and time criteria in the evaluation of a decision tree algorithm-based fuzzy rule.

Fourthly, Bayesian and k-NN classifiers in [32, 33] were evaluated only through accuracy. Fifthly, the hybrid methods proposed in [8] were evaluated using the time complexity criterion. The accuracy criterion was the only performance criterion for evaluating the hybrid methods proposed in [34–36].

Lastly, the accuracy criterion was used to evaluate the proposed methods in some works [37], including the SMIG module, [38] 2,1-norm algorithm, [39, 40] geometric algorithm, [41] algorithm of rotation forest, [42] similarity-balanced discriminant neighbourhood embedding [43] and Naive-Bayes method. In [7, 44, 45], accuracy and e-validation were employed to evaluate their methods. As for other works, [46] applied the method regression model and [47] used the ant colony optimisation model, which were based on accuracy–time and validation in performance evaluation. Elastic net for simultaneous classification was measured by error rate in [48]. The study in [49] depended on sensitivity criterion in hierarchical tree method evaluation. Conversely, [2] used time and accuracy to measure the performance of the squares regression method. The last study in this group, [50], was more comprehensive in evaluation criteria terms, as it used accuracy, specificity, sensitivity, f-measure, tp, tn, fp, fn, time

complexity, e-training and e-validation. An integrated method for classification was suggested in [51].

Numerous articles have proposed methods for feature selection/reduction. This section will focus on describing the classification performance evaluation criteria used in each study. The first three studies [3, 52, 53] used an accuracy criterion in the evaluation process. These three evaluation criteria (accuracy, specificity and sensitivity) are used in [54, 55]. Article [56] used five criteria: accuracy, tp, tn, fp and fn. Studies [57, 58] were based on accuracy, sensitivity and time complexity in the performance evaluation process. In [59], only two criteria were used, namely, accuracy and time, while [60] evaluated the performance of the proposed method according to accuracy, precision and specificity. Another work [61] used accuracy, precision, specificity, sensitivity and validation. The last four works used most of the available criteria. [62] depended on accuracy, tp, tn, fp, fn and time complexity in the evaluation section. [63] used accuracy, sensitivity specificity, tp, tp, tn and fn. [64] used accuracy, precision, specificity, sensitivity, tp, tn, fp, fn, training and validation. Finally, [65] conducted an evaluation based on accuracy, tp, tn, fp, fn, training, and validation.

The research efforts in the last group included studies on feature selection and classification. The evaluation criteria used in each study will be also described in this section. Accuracy criterion was used to evaluate the hybrid method proposed in [66]. Multiple fuzzy-rough sets were evaluated according to accuracy and sensitivity in [67]. Three evaluation criteria, namely, accuracy, training and validation, were used in [68]. Accuracy, time complexity, and training were used in [69]. These three studies [70] [71] [72] each used two criteria, namely, e-training and time complexity, accuracy and validation and accuracy and validation, respectively. A new approach that combines feature (gene) selection with transudative SVM was used in [73].

System development This category includes 22 papers that focused on providing proposals for developing the classification systems of acute leukaemia. The studies in this section are described by concentrating on the evaluation criteria used in various classification systems based on supervised or unsupervised classification engines. Two types of efforts are included in this section: 1) the evaluation criteria used in systems based on a supervised classification engine and 2) the said criteria used in systems on unsupervised classification methods. The first five studies emphasised the detection or classification system according to three phases, namely, segmentation, feature extraction and classification. However, they used different criteria. Accuracy, precision, specificity and f-measure were used in [13]; accuracy and validation were used in [74]; accuracy and time were used in [9]; and accuracy and training were used in [75, 76]. The next four studies involved four phases: preprocessing, image segmentation, extraction and/or selection of the features and data classification. Accuracy, precision, specificity and sensitivity were used for the evaluation process in [11, 15, 16], while accuracy and precision, specificity, sensitivity, tp, tn, fp and fn were represented in the evaluation criteria in [77]. Accuracy, specificity, sensitivity and time complexity were used in [21]. Accuracy, precision, specificity, and f-measure were employed in [14]. Another two works [17, 18] focused on developing classification systems that encompass six phases and used accuracy as the performance measurement. [19], which also depended on accuracy, whereas [78] used time complexity in its evaluation section. [79] used accuracy and recall as measurements. Two key phases with 10 sub-phases based on game theory were used in [6], which then evaluated the resulting method using accuracy, time complexity and validation. In [80], accuracy was used to evaluate the proposed automated detection schema of lymphoblasts. Conversely, in [81], accuracy, sensitivity, training and validation were the main evaluation criteria for the gene selection and classification system. Another research effort proposed a system of classification based on the fuzzy rule concept [82] with pre-processing, fuzzy clustering and selection, rule extraction and classification and use validation in the evaluation process. A fuzzy expert system was evaluated according to accuracy in [83]. Finally, an intelligent multi-agent was used to assist in understanding the process of classification in [1].

Evaluation and comparative study This research direction contains five articles that attempted to evaluate the classification methods or compare them according to selected evaluation criteria. The first study [84] depended on error rate criteria to compare two classification schemes, while in [10], accuracy and precision, specificity, sensitivity, f-measure, tp, tn, fp and fn were the main evaluation criteria. The remaining three [12, 20, 85] used accuracy as a main criterion.

Evaluation and benchmarking for acute leukaemia classification systems

This section describes the evaluation and benchmarking. It includes the criteria of evaluation and presents the various evaluation and benchmarking tools and their limitations. A summary of evaluation and benchmarking challenges and open issues is also reported.

Evaluation criteria

This section presents the different evaluation criteria for acute leukaemia classification tasks. These criteria were divided into two main groups: the reliability group and time complexity. Each of these groups has a subgroup.

Reliability group The reliability group includes four sub-groups of criteria (Fig. 4): the matrix of parameters, relationship of parameters [i.e. Ave Accuracy, Precision $_{(Micro)}$, Precision $_{(Macro)}$, Recall $_{(Micro)}$ and Recall $_{(Macro)}$], the behaviour of parameters [i.e. F-score $_{(Micro)}$ and F-score $_{(Macro)}$] and error rate. This section describes in detail the evaluation criteria in each group.

Matrix of parameters The matrix of parameters is the main sub-category of the reliability group. This matrix is also known as the confusion matrix. It includes the key parameters of machine learning outputs. This matrix is also commonly used in the machine learning domain [36, 86]. It is widely utilised in describing the performance of classification models [20]. The values in a confusion matrix show the predicted and actual classification class achieved by the classification system [87]. The confusion matrix describes the incorrect and correct predictions in comparison to the real results of the test samples [28]. This approach allows for a more detailed analysis than the mere proportion of correct classifications.

A confusion matrix consists of two aspects. The first dimension represents the actual classification class of an object, and the other pertains to the classification class which the classification model predicts; moreover, each cell has the corresponding number of predictions achieved by the classification model that falls into that cell [60] [64]. The size of the confusion matrix is $N \times N$, where N is the number of the various values of a label [11]. For a binary classification problem, the confusion matrix has two actual classes and two predicted classes, which mean the confusion matrix table has two rows and two columns [65]. For multiclass classification, the confusion matrix has more than two actual classes and predicted classes [36]. Figure 5 illustrates the binary classification task confusion matrix.

A confusion matrix for binary classification is formed from the four predicted outcomes (true positives, true negatives, false positives and false negatives), which are produced from the binary classification model [60]. The confusion matrix values are listed and described in Table 1.

Two of those parameters reflect the correct classification (true positive and true negative), while the false positive and false negative outcomes are two possible types of errors. Hence, this matrix allows for the identification of points which are correctly and incorrectly classified. It illustrates the performance of a classification model by displaying the actual and predicted points.

The confusion matrix is used in many of the reviewed studies to describe the performance of a binary classification model. [28] used a confusion matrix to summarise the testing results of the classification model which aimed to classify acute leukaemia into ALL and acute myelogenous leukaemia (AML). Among the test samples, 16 were in ALL and 8 were in AML. Figure 6 depicts the results of the resulting confusion matrix for binary classification. [60, 61, 64] also used confusion matrices to present the performance results of classifying acute leukaemia into AML and ALL. [11] used the confusion matrix for the results of classifying acute leukaemia into normal cells and blast cells.

The confusion matrix is also used in the performance evaluation of more general cases of multiclass classification. Multiclass evaluation is an extension of the methods used in binary evaluation, wherein multiple classes are involved instead of only two. The confusion matrix can generate multiclass cases [86]. For a multiclass classification problem, the confusion matrix is built with L classes, with L being more than two classes (Fig. 7).

In Fig. 7, the points in the grids with matching actual and predicted classes are the correct predictions. The grey grids are the grids for the correctly classified points (correct decisions made). In an ideal scenario, all other grids should have zero points or, in case of misclassifications, the values of other grids are the errors in the confusion between the various classes.

Fan et al. describes the detailed classification performance per class in the confusion matrix, and the sub-types of ALL (BCR-ABL, E2A-PBX1, MLL, T-ALL, TEL-AML1 and Others) were the classes that were used in the confusion matrix [36]. [65] used a multiclass classification technique to classify the data samples into five categories (ALL, AML, CLL, CML and Normal). Figure 8 depicts the results for multiclassification in the confusion matrix table.

In Fig. 8, the multiclass classification model successfully classified all samples correctly.

Finally, the confusion matrix describes all the results of multiclass classification tasks, and thus provides details about the correct and incorrect predictions. The parameters of the confusion matrix are the basis for evaluating the classification models, and through these parameters the rest of the evaluation metrics are calculated.

Relationship of parameters This group of evaluation criteria includes five metrics typically used to measure the quality ratio of any multiclass classification model.

- Accuracy

Classification accuracy is considered one of the most important metrics of evaluation. Accuracy expresses the performance or significance of the algorithms' behaviour [30]. Classification accuracy is a commonly used metric for evaluating the quality of a classification system [59]. The value of classification accuracy changes depending on the selected datasets [30, 86].

Accuracy is measured for multiclass classification tasks in the macro level based on confusion matrix results as follows [86, 88]:

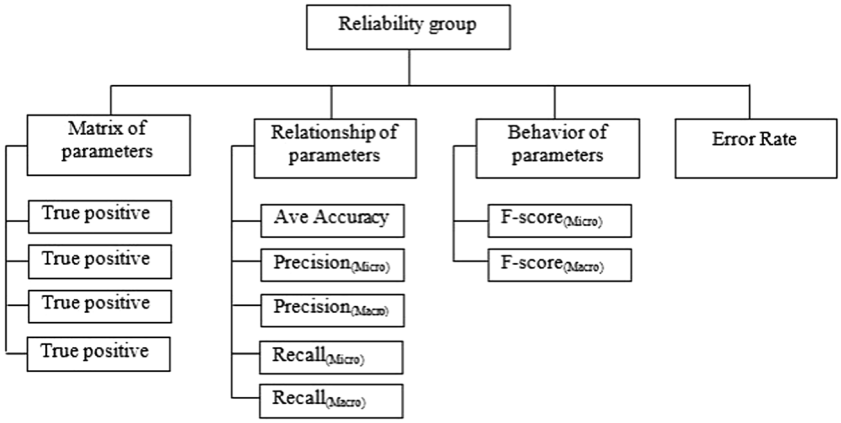

Fig. 4 Reliability group of criteria

$$Ave\ Accuracy = \frac{\sum_{i=1}^{J} \frac{tp_i + tn_i}{tp_i + fn_i + fp_i + tn_i}}{\mathrm{J}} \quad (1)$$

where Ci has many classes, for which tp_i is the true positive, fp_i is the false positive, fn_i is the false negative, and tn_i is the true negative.

Many studies have relied widely on accuracy in the performance evaluation of classification tasks, but they did not consider the differences between the types of classified classes; in real cases, especially in medicine, the distinctions between certain classified classes are very important [76].

- Precision

Accuracy can be measured provided that a specific class has been predicted [60]. Precision (P) is defined as the sum of true positives (TP) over the sum of true positives plus the sum of false positives [16]. Precision is measured for multiclass classification tasks in the micro and macro levels according to the results of the confusion matrix [86, 88]. Table 2 describes the equations for each precision level that can be used.

Precision provides the ratio of subjects with positive outcomes that are correctly identified [60]. Thus, precision measures the classifier exactness. A low number of false positives means high precision, and vice versa [87].

- Recall

Recall is also called sensitivity or rate of true positive, and it indicates the test's capability to determine positive outcomes [16]. It is a measure of a classification model's capacity to identify the instances of a specific class from a dataset [87].

		Actual class	
		Class1	Class 2
Predicted class	Class 1	True positive (TP)	False negative (FN)
	Class 2	False positive (FP)	True negative (TN)

Fig. 5 Binary classification confusion matrix

Recall depicts the completeness of a classification model. A low number of false negatives means a high recall, whereas a high number of false negatives means low recall [87]. The higher the recall, the better the classification model becomes. It determines the ratio of accurately classified samples to total samples [50, 60, 89].

Recall is measured for multiclass classification tasks in the micro and macro levels according to the results of the confusion matrix [86, 88]. Table 3 describes the equations for each precision level that can be used.

Behaviour of parameters This section describes two measures, namely, $\mathrm{Fscore}_{(micro)}$ and $\mathrm{Fscore}_{(macro)}$. Fscore refers to the harmonic mean of recall and precision. It gives the overall performance of a classification model [14, 16, 60]. We can examine the combined performance through this metric [50].

Fscore is measured for multiclass classification tasks in the micro and macro levels according to the results of the confusion matrix [86, 89]. Table 4 describes the equations for each precision level that can be used.

Error rate within the dataset group This criterion measures errors made by the classification model. It is one of the main criteria in the evaluation and benchmarking of classification systems for acute leukaemia [1, 60, 66]. Roy et al. stated that determining the best classification model depends on the results of the error rate, and among some available classifiers, they chose the one that achieved the best outcome according to the error rate metric on the training and validation sets [72]. The lower the error rate, the better the classification model [64, 74]. [40] indicated that performing cancer prediction with a small error rate requires a comparatively big training sample set in the classification model learning. The importance of the error rate as a criterion in evaluation and benchmarking is supported by Shi et al., who used error rate in the comparison of their method with previous techniques [61]. The evaluation of the error rate of a classification model using the training dataset is considered an unreliable criterion. In fact, the classification model can overfit the training dataset with such a

Table 1 Confusion matrix parameters

Parameters		Description
True Positives	TP	The number of elements correctly classified as positive by the test [10, 28]. When cancer cells are correctly identified [15].
False Positives	FP	The number of elements classified as positive by the test, but they are not [10, 28]. when non-cancer cells are identified as cancerous [15].
True Negative	TN	The number of elements correctly classified as negative by the test [10, 28]. when non-cancer cells are correctly identified [15].
False Negatives	FN	The number of elements classified as negative by the test, but they are not [10, 28]. when cancer cells are identified as noncancerous [15].

strategy. Thus, we need to use an independent dataset that differs from the one used in the training process (test data sample) to measure the error rate of the classification model. Hence, the data sample is divided into the number of parts using N-fold cross validation. The said validation is a random division of the sample of the dataset into N parts. Then, one of those parts is used for testing the classification model, and the others parts are used for the classification model learning [66].

Error rate is measured for multiclass classification tasks in the macro level according to the results of the confusion matrix using the following equation [86, 89]:

$$\text{Error rate} = \frac{\frac{\sum_{I=1}^{J} tp_i + tn_i}{\sum_{i=1}^{J} tp_i + fn_i + fp_i + tn_i}}{J} \quad (8)$$

Table 5 presents the survey of the reliability group criteria used in various reviewed studies.

Table 5 provides a comprehensive review of the different evaluation metrics in evaluating and benchmarking the acute leukaemia detection and classification in all reviewed studies. The largest ratio of studies (90%) used accuracy. The other percentages are as follows: error rate (21%), recall (26%) and specificity (30%). The ratio of usage of other metrics with less than 20% are as follows: FP% (16%), FN% (13%), precision (15%), TP% (15%), TN% (3%) and fscore (1%). Variations occurred in the percentage of usage for these metrics among various reviewed studies and conflict was noted among the sub-metrics (Table 4). Additionally, no study used all this set of metrics together. The usage variance of these metrics indicates a serious challenge in using a specific set of metrics when evaluating and benchmarking the detection and classification of acute leukaemia classification. Such varying usage rates of evaluation metrics also suggest that no common guideline exists for evaluating various metrics and that each study applied the metrics that fulfils its objectives.

		Target	
		ALL	AML
Output	ALL	16	0
	AML	0	8

Fig. 6 Sample confusion matrix for binary classification

Time complexity Time complexity is one of the challenges faced by researchers seeking to develop an acute leukaemia classification system. One of the main requirements for acute leukaemia classification is obtaining the least time complexity [55, 57, 72]. Time complexity is a significant metric in the evaluation and benchmarking of classification models [67]. [90] stated that one of the important requirements of acute leukaemia classification is rapid detection and classification, and the system which consumes less time, especially for medical images, can help save lives through the early detection of disease and prompt treatment. Time complexity is the time consumed by the input and output of sample images, which means the time required to complete the classification task of that algorithm [91]. One of the disadvantages of a classifier is the time-consuming processing [90, 91]. Time complexity is also vital in evaluating the performance efficiency of a system based on image processing; the best classifier is the one that achieves the highest accuracy and the least time complexity [77]. Processing time depends on the size of the dataset [70]. Table 6 illustrates the survey of time complexity criteria used in various reviewed studies.

Table 6 indicates that 24% of the reviewed studies applied this metric to measure the processing time. All remaining studies did not mention this metric. Processing time is one of the main requirements that should be measured in the classification systems of acute leukaemia. However, many of the reviewed studies focused on other metrics and overlooked time complexity.

Benchmarking tools

This section aims to review the various tools used in benchmarking different classification systems or methods. These tools compare the outputs of various classification systems according to specific criteria. The purpose of this comparison is to ascertain the quality enhancement of a new system against previous or other approaches. These tools are developed mainly to fulfil machine learning processes and activities, as well as data analysis [92], but they have also been used in many studies for benchmarking processes [8, 20, 25, 31, 93], However, benchmarking based on these tools indicate incomplete benchmarking processes.

Fig. 7 Confusion matrix for multiclass classification

		Actual class				
		Class 1	Class 2	Class 3	…	Class n
Predicted class	Class 1	Correct (TP_1)				
	Class 2		Correct (TP_2)			
	Class 3			Correct (TP_3)		
	…				----	
	Class n					Correct (TP_n)

Common benchmarking tools

Many tools can used in benchmarking. The widely used examples are described in the following sections.

Rapid miner Rapid Miner is a software application that provides many tools and function libraries for machine learning applications. Developed by the Rapid Miner Company, it offers a comfortable and friendly environment based on a client/server model with two server options as service or as cloud. It can be run across different platforms and includes many machine learning methods and more than 100 schemes. Its methods support various analyses, such as clustering, classification and regression. In addition, it allows for flexibility regarding file formats, as it can accommodate about 22 formats. It was first developed in 2001 at the Technical University of Dortmund and was known then as YALE. Its name was later changed to Rapid Miner [93–96].

Weka Weka is an integrated platform containing a group of methods, algorithms and visualisation possibilities for the modelling and analysis of data. It provides friendly GUI for ease of use. It supports many machine learning and data mining applications, such as association, clustering, classification and preprocessing. It was first developed in the Waikato University labs. It runs on Linux, Windows, and OS X systems. Weka was developed initially for the analysis of agricultural data. The new version based on java was first developed in 1997 and is applicable to different domains [3, 20, 31, 62, 94].

R tool R software is a free tool that provides an integrated statistical and visualisation environment. The R Core Team developed it, and the first version was released in 1997. It supports work in multi platforms. S language was used to develop the R tool, which was deployed as open source software. Many extra packages were developed and provided as a free and publicly available resource. It supports various applications, such as statistical and data mining applications. It provides advanced and complex statistical tools and includes most formulae and mathematical symbols. This tool is considered easy to use and applicable for sampling [94–96].

Konstanz information miner Konstanz Information Miner (KNIME) is an integration environment for data analysis and reporting. It was developed by KNIME AG to support various ML and DM algorithms and activities. The first version of KNIME was created in 2004 at the Konstanz University by software engineers. It works on Windows, Linux and OS X systems. It includes more than 100 methods for data cleaning and preprocessing, data mining, cleansing and data analysis and also provides different interactive visions for parallel coordinates, scatter plots and others [94, 95].

Orange The Orange tool is one of the commonly used analytical software. It covers a wide range of DM algorithms which perform many activities, such as scoring of features and filtering. Python and C++ were the main languages used to develop this tool, so it is characterised by flexibility and robustness [95].

KEEL KEEL is a data mining tool that includes many traditional methods and techniques for data processing and knowledge extraction. It can deal with different file formats (e.g. ARFF, CSV and XML) and provides possibilities for learning depending on intelligence and building simple, hybrid models and statistical modules. It also covers different feature processing, including selection and discretisation [96] .

		Actual classes				
		ALL	AML	CLL	CML	Normal
Predict classes	ALL	8	0	0	0	0
	AML	0	8	0	0	0
	CLL	0	0	8	0	0
	CML	0	0	0	8	0
	Normal	0	0	0	0	8

Fig. 8 Sample of confusion matrix for multiclass classification

Weaknesses of the reviewed benchmarking tools

Although the tools described in the previous section are widely used in the machine learning field [93], they suffer from many weaknesses in the process of evaluation and benchmarking.

Not all the tools described above were used to compare acute leukaemia classification systems. Moreover, the tools used for the evaluation and benchmarking process were utilised according to some, but not all, the evaluation criteria. Hence, such usage does not reflect all the necessary aspects of evaluation and benchmarking. The weaknesses of current tools in terms of evaluation and benchmarking include [97–99] failure to calculate the overall parameters of the

Table 2 Precision metrics of multiclass classification tasks

Precision $_{\mu}$	$\frac{\sum_{i=1}^{J} tp_i}{\sum_{i=1}^{J}(tp_i+fn_i)}$ (2)	Agreement of the data class labels with those of classifiers if calculated from sums of per decisions
Precision $_{M}$	$\frac{\frac{\sum_{i=1}^{J} tp_i}{\sum_{i=1}^{J}(tp_i+fn_i)}}{J}$ (3)	An average per-class agreement of the data class labels with those of classifiers

Where Ci many classes, *tpi* are true positive for Ci, and *fni*- false negative, counts respectively. μ and *M* indices represent micro- and macro- averaging

reliability group of metrics, inability to compare among multiclass classification models simultaneously using more than one criteria, failure to match multiclass classification models and inability to rank multiclass classification models from best to worst.

Thus, a new method for evaluation and benchmarking that covers all important and required aspects is necessary. Including all the measurement criteria (reliability group and time complexity) during the evaluation and benchmarking will generate accurate results that reflect the quality of all aspects of the multiclass classification models of acute leukaemia. In turn, such accuracy would allow us to choose the appropriate classification model according to its evaluation and compare it with other models from multiple aspects and on the basis of different criteria.

Discussion

This section presents a detailed description and analysis of the challenge and open issues related to evaluation and benchmarking, in-depth analysis of the studies that attempted evaluation and benchmarking and an explanation of the automated multiclass classification of acute leukaemia. It also includes the recommended solution. The review of prior studies highlighted three serious open issues resulting in the difficulty of evaluation and benchmarking of the multiclass classification of acute leukaemia: conflicting criteria issue, criteria importance and evaluation of criteria issues. Multicriteria decision analysis is proposed as a solution. A decision support system based on MCDM is suggested for the evaluation and benchmarking of the automated multiclass classification of acute leukaemia. The recommended decision support system has three sequential phases. Phase One presents the identification procedure and the process for establishing a decision matrix based on a crossover of evaluation criteria and acute leukaemia multiclass classification models. Phase Two describes the decision matrix development for the selection of acute leukaemia classification models according to integrated BWM and VIKOR. Phase Three involves the validation of the proposed system.

Challenge and open issues related to evaluation and benchmarking

Recently, evaluation and benchmarking classification tasks associated with critical medical fields have gained growing interest to obtain high-performance classification processes. Such efforts confront problems and issues in several significant aspects. The issues and challenges are found in the evaluation and benchmarking of acute leukaemia multiclass classification systems. Benchmarking is carried out after the development of any system, with the aim of comparing the new system with other similar systems under the same conditions and metrics [9]. Evaluation and benchmarking involve verifying whether the newly developed multiclass classification systems satisfy the requirements. The main requirements for automated multiclass classification systems for acute leukaemia are high reliability versus decreased time complexity and high accuracy versus low error rate [90]. Achieving these requirements simultaneously poses a challenge [47]. Given the difficulty in overcoming this challenge, we find that most studies focused on one requirement and neglected the rest. This situation causes a conflict between criteria during the comparison process. Such conflict will be reflected in the evaluation and benchmarking. As a result, the benchmarking process is affected because benchmarking between multiple conflicting criteria is problematic [83]. In addition, the current comparison approach between the proposed systems and the previous systems in all the reviewed studies does not consider all evaluation and benchmarking criteria and instead concentrates on one aspect of the evaluation, overlooking the rest because it is not flexible enough to address the conflict between the various

Table 3 Recall metrics of multiclass classification tasks

Recall $_{\mu}$	$\frac{\sum_{i=1}^{J} tp_i}{\sum_{i=1}^{J}(tp_i+fp_i)}$ (4)	Effectiveness of a classifier to identify class labels if calculated from sums of per decisions
Recall $_{M}$	$\frac{\frac{\sum_{i=1}^{J} tp_i}{\sum_{i=1}^{J}(tp_i+fp_i)}}{J}$ (5)	An average per-class agreement of the data class labels with those of classifiers

Where Ci many classes, *tpi* are true positive for Ci, and *fpi*- false positive, counts respectively. μ and *M* indices represent micro- and macro- averaging

Table 4 Fscore metrics of multiclass classification tasks

F-score$_{\mu}$	$\frac{(\beta^2+1)precision_{\mu}Recall_{\mu}}{\beta^2 precision+_{\mu}Recall_{\mu}}$ (6)	Relations between data's positive labels and those given by a classifier based on sums of per decisions.
F-score$_{M}$	$\frac{(\beta^2+1)precision_{M}Recall_{M}}{\beta^2 precision_{M}+Recall_{M}}$ (7)	Relations between data's positive labels and those given by a classifier based on a per-class average.

Where Ci many classes, *tpi* are true positive for CI, and *fpi* – false positive, counts respectively. μ and *M* indices represent micro and miacro- average

criteria. Therefore, multiclass classification tasks require better evaluation and benchmarking. The following subsections will explain the main issues that cause the challenge in evaluation and benchmarking. Figure 9 illustrates the main issues of evaluation and benchmarking in automated acute leukaemia classification.

Conflicting criteria issue An important issue found in related literature is the conflict or tradeoff between different performance criteria. The tradeoff situation results in the loss of one or more aspects of the performance quality of acute leukaemia classification systems. Tradeoff requires the users to give up one requirement for another owing to the difficulty of achieving balance between all requirements. In our case, a conflict occurs among criteria that measure the basic requirements of any acute leukaemia classification system [21]. These criteria are related to identifying the strengths and weaknesses of each system, which leads to the inability to make a rational decision for evaluation and benchmarking the different alternatives and selecting the best one among them. The varying ratios among the different criteria collected in our study also showed the effect of the conflict on various criteria used by researchers. Consequently, the conflict amongst evaluation criteria for acute leukaemia classification systems constitutes a formidable challenge in our intention to create a skin cancer segmentation/classification approach. This challenge mainly arises from conflicting terms, particularly, the conflict amongst the criteria and amongst the data. The reviewed studies demonstrate conflicting criteria or tradeoff problems between reliability, time complexity of the acute leukaemia classification model and error rate within the dataset in the evaluation and benchmarking of classification systems. Reliability should be high, time complexity for conducting the output images should be low, error rate resulting from the training datasets should be low and accuracy should be high. In the development of multiclass classification systems for acute leukaemia, all development requirements must be taken into account [90]. Conflicting data are observed due to the section matrix of parameters on TP, FP, TN and FN, which show the rise in TP and TN when parameters FP and FN are reduced [13, 28, 50, 62]. By contrast, a comprehensive assessment and benchmarking methodology covering all evaluation criteria and capable of dealing with the conflict criteria should be used to ensure a successful system that achieves its objectives [22]. This status of conflicting data pointed to an obvious conflict between the probability parameters. Such parameters significantly affect the values of the rest of the metrics within the reliability group. Thus, such requirements must be considered during evaluation and benchmarking. Each reviewed study reported that evaluation and benchmarking of all criterions are independent of the general framework. Accordingly, the approach of acute leukaemia classification must be performed to standardise basic and advanced requirements, and a clear methodology must be implemented during research for testing, evaluation and benchmarking. A new and flexible evaluation and benchmarking method must be applied to address all conflicting criteria and data problems. However, to our knowledge, solutions in this aspect have not yet been suggested on these particular issues.

Issue of criterion importance The evaluation of acute leukaemia classification systems involves a set of criteria, and the importance of each criterion varies according to the objectives for which the system is developed. In other words, the importance of one of the evaluation criteria may be increased in exchange for the low importance of another criterion according to the objectives of the system. Thus, a conflict will exist between evaluation and benchmarking criteria, due to the different importance of each criterion in different systems [47]. The conflict status among the criteria is one of the serious challenges for the evaluation process. Suitable action must be prescribed for a situation wherein the importance of a certain criterion is increased while that of others is decreased. Two key sides should be taken into account. First, the behaviour of the classification systems of acute leukaemia must be understood and achieved, thereby giving specific significance to the design. Second, the approach must be evaluated by considering the tradeoff.

The evaluator's opinions may disagree with the designers' aims as well, which can impact the final evaluation of the needed approach. Technically, evaluation and benchmarking of the classification systems of acute leukaemia entail taking into account multi criteria simultaneously, including rate of time complexity and reliability with their sub-criteria [10, 77, 87] and assigning the favourable weight for all aspects to benchmark the approaches of acute leukaemia classification. After comparing the scores of all approaches, those with the 'highest balancing rate' must receive the highest priority level, and those with the 'least balancing rate' must be given the lowest priority levels. Evaluation and benchmarking are difficult tasks and can

Table 5 Reliability group criteria used in various reviewed studies

No	References	Reliability group									
		Relation of parameters				Behaviour of Parameters	Matrix of Parameters Section				Error Rate
		Accuracy	Precision	Specificity	Sensitivity	F- score	Confision matrix				
							TP	TN	FP	FN	
1	[25]	*									
2	[70]										*
3	[44]	*									*
4	[79]	*			*						
5	[78]	*									
6	[8]	*									
7	[82]										*
8	[78]	*									
9	[81]	*			*						*
10	[75]	*									*
11	[28]	*	*	*	*		*	*	*	*	
12	[30]	*									
13	[20]	*									
14	[84]										*
15	[80]	*									
16	[6]	*									*
17	[37]	*									
18	[13]	*	*	*	*		*	*	*	*	
19	[14]	*	*	*	*	*					
20	[18]	*									
21	[53]	*									
22	[19]	*									
23	[21]	*		*	*						
24	[76]	*									
25	[7]	*									*
26	[34]	*									
27	[24]	*									*
28	[31]	*		*	*						
29	[35]	*									
30	[26]	*									*
31	[55]	*		*	*						
32	[68]	*									*
33	[60]	*	*	*							
34	[52]	*									
35	[62]	*					*	*	*	*	
36	[63]	*		*	*		*	*	*	*	
37	[64]	*	*	*	*		*	*	*	*	
38	[65]	*					*	*	*	*	*
39	[58]	*									
40	[59]	*									
41	[57]	*			*						
42	[54]	*		*	*						
43	[3]	*									
44	[2]	*									

Table 5 (continued)

No	References	Reliability group									
		Relation of parameters				Behaviour of Parameters	Matrix of Parameters Section				Error Rate
		Accuracy	Precision	Specificity	Sensitivity	F- score	Confision matrix				
							TP	TN	FP	FN	
45	[23]	*									
46	[69]	*									*
47	[46]	*									
48	[84]				*		*	*	*	*	
49	[32]	*									
50	[29]	*									*
51	[43]	*									
52	[42]	*									
53	[38]	*									
54	[47]	*									
55	[67]	*			*						
56	[31]	*		*	*						
57	[50]	*		*	*	*	*	*	*	*	*
58	[27]	*									
59	[39]	*									
60	[56]	*					*	*	*	*	
61	[33]	*									
62	[22]	*		*	*						
63	[35]	*									
64	[66]	*									*
65	[15]	*	*	*	*						
66	[87]	*	*		*		*	*	*	*	
67	[41]	*									
68	[83]	*									
69	[40]	*									
70	[74]	*									*
71	[61]	*	*	*	*						*
72	[45]	*									*
73	[48]										*
74	[71]	*									*
75	[9]	*									
76	[77]	*	*	*	*		*	*	*	*	
77	[11]	*	*	*	*						
78	[10]	*	*	*	*	*	*	*	*	*	
79	[36]	*									
80	[12]	*							*	*	
81	[16]	*	*	*	*	*					
	Frequently	75	12	19	25	3	12	12	13	13	21
	%	90	15	23	30	4	15	15	16	16	26

be regarded as extremely challenging because all classification approaches of acute leukaemia show multiple attributes that must be taken into account. For example, error rate and rate of time complexity have been proven to be very significant in the classification of acute leukaemia because they offer an objective complement to the acute

Table 6 Time complexity criteria for multiclass classification

No	References	Time complexity
1	[25]	
2	[70]	*
3	[44]	
4	[79]	
5	[78]	
6	[8]	*
7	[82]	
8	[78]	
9	[81]	
10	[75]	
11	[28]	*
12	[30]	*
13	[20]	
14	[84]	
15	[80]	
16	[6]	*
17	[37]	
18	[13]	
19	[14]	
20	[18]	
21	[53]	
22	[19]	
23	[21]	*
24	[76]	
25	[7]	
26	[34]	
27	[24]	
28	[31]	
29	[35]	
30	[26]	
31	[55]	
32	[68]	
33	[60]	
34	[52]	
35	[62]	*
36	[63]	
37	[64]	
38	[65]	
39	[58]	
40	[59]	*
41	[57]	*
42	[54]	
43	[3]	
44	[2]	*
45	[23]	*
46	[69]	
47	[46]	*
48	[84]	
49	[32]	
50	[29]	
51	[43]	
52	[42]	
53	[38]	
54	[47]	*
55	[67]	
56	[31]	*
57	[50]	*
58	[27]	*
59	[39]	
60	[56]	
61	[33]	
62	[22]	*
63	[35]	
64	[66]	
65	[15]	
66	[87]	
67	[41]	
68	[83]	*
69	[40]	
70	[74]	
71	[61]	
72	[45]	
73	[48]	
74	[71]	
75	[9]	*
76	[77]	
77	[11]	
78	[10]	
79	[36]	
80	[12]	*
81	[16]	
	Frequently	20
	%	24%

leukaemia classification decision and optimise inter-rater consistency. Consequently, for these attributes, different weights may be provided by each decision maker. On the one hand, developers who aim to give a score for an acute leukaemia classification approach might assign more weight to one feature rather than to other features that attract less interest. On the other hand, developers who aim to use benchmarking software to solve such problems will probably target various attributes as the most significant one, such as the accuracy [2, 23, 69]. Thus, evaluation and benchmarking for classification approaches of acute leukaemia suffer from highly complex attribute problems.

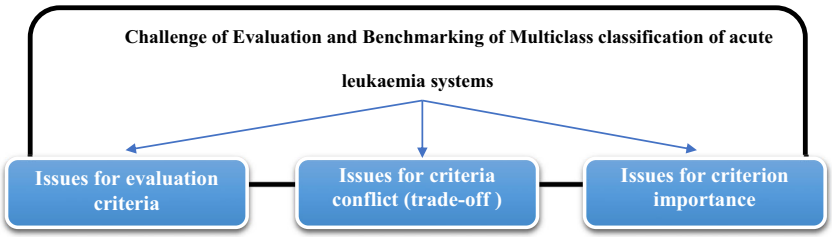

Fig. 9 Main issues in the evaluation and benchmarking of automated acute leukaemia classification systems

Issue for the evaluation criteria Numerous critiques have been performed on the criteria of evaluation. A problematic figure exists on the variation of error rate values in dataset criticism resulting from the varying sizes of the datasets used in different acute leukaemia experiments [61, 68]. Thus, one important issue of these criteria must consider the error rate value with each experiment owing to the lack of a standard dataset; in addition, an unjustified consumption of effort and time exists caused by an unorganised collection of dataset, depending on individual studies [10]. The reason for criticising the reliability set of the criteria is that its result depends on the confusion matrix that contains four parameters, namely, TP, FP, TN and FN. The number of pixels may be lost during the cropping of the image background of acute leukaemia using an image editor when manually labelling the actual class; moreover, comparing the actual class with the predicted class to compute one of the matrices of parameters are needed [10, 18, 79, 84]. Thus, this status will affect the results from all reliability sets (behaviour, matrix and relationship) of parameters, which are considered debatable. Although the critiques for these criteria exist in many earlier studies, they are still extensively used for evaluating the various tasks of acute leukaemia detection and classification.

Critical review and analysis

The growing number of available classification models and systems of acute leukaemia is considered a major problem for health organisations and other cancer treatment centres. Health organisations specialising in cancer treatment have encountered a challenge on how to select the appropriate acute leukaemia classification system that would allow accurate and rapid detection and classification of acute leukaemia. Previous studies have clearly demonstrated that the acute leukaemia classification tasks vary in terms of the accuracy of the results they provide, apart from the overall performance disparities. In the same context [15, 16, 20], no single classification system is confirmed to be superior over the rest. According to our systematic survey on automated classification of acute leukaemia, a total of 83 studies addressed the different aspects of such a classification system. The explanation in 'Section 2.2.1. Proposed Methods' demonstrates that most of the reviewed studies attempted to propose developing a new classification method; alternately, other studies attempted to enhance the current methods. Most of those studies were focused on enhancing the accuracy of classification, decreasing the classification time or improving the overall performance of classification. In addition, the analysis in 'Section 2.2.2. System development' demonstrates that numerous studies have developed classification systems of acute leukaemia, and these systems differ in terms of classification techniques, phases and procedures of classification; different accuracies of the classification results are also provided by each system. The analyses in Sections "Evaluation criteria". and 2.2.2. presented many of the methods; models or systems of classification of acute leukaemia have been proposed or developed in the literature. Obviously, all those works differ from one another in terms of classification techniques used and procedures followed, as well as the difference in the accuracy of the classification results from one to the other. Apart from the disparity of their overall performance, all results confirm the difficulty of making a decision to choose a better option among them.

However, no study has provided a comprehensive and integrated solution to assist in evaluating and benchmarking the multiclass classification models or systems to determine a suitable one. In 'Section 2.2.3. Comparative and Evaluation Study', five studies were described in Table 7, which attempted to address the evaluation and benchmarking issues. However, our analysis of their work found that they only attempted to evaluate the classification tasks of acute leukaemia on the basis of partial dimensions. Alternately, they compared among several models of acute leukaemia classification on the basis of individual criteria. Therefore, their solution cannot be used as basis for taking a complete picture that reflects all dimensions of evaluation and comparison to choose the right decision for a suitable solution of acute leukaemia classification.

As shown in Table 7, studies that focused on the evaluation and comparison of acute leukaemia dealt with a few aspects of this evaluation and neglected other aspects. By contrast, [12, 20] depended on the evaluation and benchmarking of the accuracy aspect only. At the same time, Snousy et al. confirmed that accuracy is an important criterion in cancer classification task but is not the only goal in the cancer domain; their study aimed to investigate the effectiveness of various features of the selection methods on classification accuracy among different classification models. [10, 85] attempted to evaluate and benchmark on the basis of a few of the reliability group criteria. Rota et al. depended on accuracy, precision and recall; whereas Labati et al. focused on TP, TN, FP, FN, specificity,

Table 7 Literature survey of various studies in the evaluation and benchmarking of automated classification tasks for acute leukaemia classification

Author & year	Brief Description	Used Criteria
[85]	This study proposes a comparison among three different approaches for the automatic detection of leukemic cells. The first based on support vector machine, the second based on neural network, and the third based on gaussian mixture model estimation and bayes decision.	Accuracy, precision, recall
[20]	This study provides a performance comparison of nine classifications models based on decision tree techniques. It attempted to experimental exams effect to different features selection methods on classification accuracy.	Accuracy
[84]	This study provides a performance comparison of two classification schemes with respect to the segmentation quality and effect different segmentation on classification results. The first scheme based on support vector machines and the second based on random forests.	Error rate
[10]	This study proposes ALL-IDB, a public image dataset of peripheral blood samples of normal individuals and leukemic patients, which provides a supervised classification and segmentation of the data., specifically designed for the evaluation and the comparison of algorithms for segmentation and classification.	TP, TN, FP, FN, specificity, sensitivity and accuracy
[12]	This study aims to analyze the performance of automated microscopy with DM96TM and we studied its ability to correctly identify blood cells and accuracy compared with manual method and/or XE-2100TM.	Accuracy

sensitivity and accuracy. Finally, [84] dealt with the evaluation and comparison regarding error rate. To make a substantive judgment on the quality and performance of acute leukaemia classification systems, an evaluation and benchmarking method are required which covers all the main requirements and cannot be assessed from only single aspect. In the same context, Saritha et al. confirmed that the automated classification system should have high accuracy and efficacy, less processing time, small error and robust. Early identification of leukaemia yields in providing the appropriate treatment to the patient [90]. This study attempts to fill the gap in the evaluation and benchmarking of the acute leukaemia classification area. This study provides a new decision support system for the evaluation and benchmarking of multiclass classification of acute leukaemia that includes all the key evaluation and benchmarking metrics. This system shall be capable of assisting the administrations of health organisations and various users to evaluate and benchmark acute leukaemia multiclass classification solution. It can also ensure that the selected classification models meet all necessary requirements.

Automated classification of acute leukaemia

Acute leukaemia

Blood is an essential component in the human body, and it achieves many of the important functions related to maintaining the metabolism process by delivering oxygen and other vital minerals. White blood cells (WBC), red blood cells (RBC) and platelets are the essential components of blood [32]. Blood cancer is one of the most serious types of cancer [19]. Leukaemia is a type of blood cancer distinguished by an irregular or abnormal growth in the number of white cells in the blood known to be immature blasts [16, 37]. Two kinds of abnormal white cells can turn into leukaemia, namely, myeloid and lymphoid cells [90]. [14, 18, 90] described acute and chronic leukaemia as two main types of leukaemia; the first type is characterised by a quick progression unlike the second type, which grows slowly. In an acute leukaemia case, irregular white cells called immature blasts work improperly. The immature blasts increase quickly and will worsen if not controlled immediately. By contrast, in the chronic type, young blood cells are present, but only the mature ones produce functional cells. Acute leukaemia results from a rapidly increasing production of white blood cells which then results in an abnormal increase in irregular cells or for the latter to be passed into the blood stream [32, 100]. Acute leukaemia starts in the bone marrow and blood and rapidly progresses. Abnormal white cells can grow in adults and children [29]. Based on the famous model of leukaemia categorisation, acute leukaemia is divided into two types, namely, acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML) [14, 53, 79]. Each type includes a number of subtypes; ALL has two subtypes (B cell ALL and T cell ALL), while AML is categorised into eight subtypes (M0, M1, M2, M3, M4, M5, M6 and M7) [79]. ALL comprises rapidly increasing infected lymphocytes. Lymphocytes are a type of white blood cells which fight infection [11]. ALL will be fatal if no remedial action is taken immediately because it is characterised by a rapid proliferation into the different body organs, especially blood circulation. Therefore, a rapid diagnosis of this type of acute leukaemia is important for the patient's recovery [13, 16]. In AML, these immature cells do not develop and are incapable of warding off infections [101]. ALL usually occurs in young children, whereas AML infects adults more than children [14, 102]. The diagnosis of acute leukaemia is mostly difficult and requires quick treatment; thus, this type requires

modern and non-traditional methods to assist the physician in making a diagnosis [15, 81, 102].

Acute leukaemia classification

Data classification is a generally common topic in various sciences, such as statistics, computer science and decision science [185]. It has many applications in medicine, engineering and management and can address many issues related to data recognition, diagnosis and detection, among others [6]. The accurate and fast detection of acute leukaemia and its classification into its subtypes are important when making appropriate remedial action [7, 8, 52]. Diagnosis and classification of acute leukaemia is a field that requires automation, especially as the manual methods for detection and classification suffer from several limitations; conversely, satisfying the need for rapid and accurate methods that can facilitate early detection and prediction of cancerous patterns [1–3, 80]. Automated classification is extensively applied to analyse cancer and is relatively a fledgling and an interdisciplinary technology that integrates the primary ideas of digital image processing [101].

Many studies have been conducted to automatically classify acute leukaemia and their subtypes and thus enhance early diagnosis [19]. An automatic classification of the ALL, images of blood or bone marrow are processed using image processing techniques **[11]**. A well-planned classification system plays a significant role in the accurate classification of acute leukaemia. Classification systems not only help experts make the right decision but also minimise possible errors [6]. A computerised system will be helpful for the analysis of stained microscopic images of the blood cells [13]. In practice, an identity from one of the known classes is utilised to assign the unknown test parameters; this task is known as classification. The classifier uses the set of features and identifies the difference between normal and diseased cells [77]. Classification is the task of associating the appropriate class label with the blood test sample by using the measurements [100].

Classification models classify a cell as normal cell or cancer-affected cell, that is, a blast cell. Alternately, they classify the cell into subtypes by comparing a few of the features [90]. Choosing the most appropriate classification methods is essential to improve classification performance [81].

When leukaemia is classified into two classes only, namely, normal and cancer cells (abnormal) or AML and ALL, this classification task is called a binary classification; whereas when the disease is classified into more than two classes, namely, normal cell, AM and ALL or, L1, L2 and L3, this classification task is called multiclass classification [8] [103].

Many machine learning methods can be employed for the multiclass classification of acute leukaemia into subtypes. The most employed classification methods will be described in the following subsections:

- **Artificial Neural Network (ANN)** is one of most popular methods among the artificial intelligence fields. Numerous authors mentioned that ANN has significant capability in interpreting and analysing medical data sets. It can represent complex patterns depending on a mathematical model that works in a way that simulates the human brain [32] [187].
- **Support Vector Machine (SVM)** is considered one of the oldest and most important methods of artificial intelligence. According to earlier studies, SVM is the most extensively used in the classification of acute leukaemia. It executes the procedures of classification by building hyper-planes in a multidimensional space that distributes cases of dissimilar and different class labels [102, 104] [186].
- **Decision Tree (DT)** is a classical model of machine learning. The structure of this method is similar to a tree; each of its branches represents a class of sample with similar characteristics. Many decision tree extensions have been developed, such as CART, C4.5, ID3 and EG2 [41].
- **Nearest Neighbour (NN)** is one of the commonly used classification algorithms. It works under supervised and nonparametric approach [30]. Classification in NN is performed by the votes of nearby neighbours. Depending on the principle of voting, objects will be identified according to their related classes [104].
- **Random Forest (RF)** [41] is an ensemble machine learning method that contains a number of DTs through a random division of the feature space. It is constructed by collecting multiple DTs and works under supervised and nonparametric approach. The principle of its work is to segment the feature space into a number of subspaces and extract the most significant features; this process is repeated until the most distinguishable training dataset and the basic classification method for various feature subspaces are obtained [104].
- **Bayesian Network** is a joint distribution based on the probability for a group of random variables to have a potential mutual causal relationship. In this method, the variables are represented by nodes; the causal relationship between each node pair is node edge and a conditional probability distribution in each of the nodes [33] [187].

Recommended pathway solutions for future direction

The previous sections described the existing evaluation and benchmarking approaches. Those sections described the evaluation methods, evaluation criteria and the challenges and issues of acute leukaemia multiclass classification evaluation. This section will present the new recommendation pathway solution. The supporting reviews are presented as follows.

The processes of evaluation and benchmarking of multiclass classification for acute leukaemia involve considering simultaneous multiple attributes (time complexity rate, reliability). Thus, a decision support system is proposed based on the MCDM method to solve multiple criteria attributes that may increase the quality of decision making [205, 206] [193]. In the real world, beneficial methods that address MCDM issues are introduced as the recommended solutions that support decision makers in solving the problems and performing analyses, evaluation and ranking [105] [192, 207].

Multi-Criteria decision making: definition and importance

Keeney and Raiffa [106] define multi-criteria decision making (MCDM) as 'an extension of decision theory that covers any decision with multiple objectives. A methodology for assessing alternatives on individual, often conflicting criteria, and combining them into one overall appraisal…' In addition, Belton and Stewart [107] define MCDM as 'an umbrella term to describe a collection of formal approaches, which seek to take explicit account of multiple criteria in helping individuals or groups explore decisions that matter'. MCDM is one of the most well-known techniques for decision making and deals with the complex decision problems in handling multiple criteria [108, 109] [191]. It presents a systematic method of solving the decision problems on the basis of multiple criteria [109]. The aim is to assist decision makers in dealing with such problems [110]. The methods and procedures of MCDM frequently depend on quantitative and qualitative approaches, and it is often focused on simultaneously dealing with multiple and conflicting criteria [111, 112] [207, 208]. Depending on the approaches, MCDM can increase the decision quality through effective and rational methods more than the traditional processes [113]. MCDM aims to attain the following: (1) categorise the suitable alternatives among a group of available alternatives, (2) rank the suitable alternatives on the basis of their performance in decreasing order and (3) select the best alternative [105] [205, 206]. Based on these goals, the suitable alternative(s) will be scored. The essential terms requiring definitions in any MCDM solution, namely, the evaluation matrix or decision matrix, are also the decision criteria [114]. A decision matrix consists of n criteria and m alternatives that must be created. The intersection of each criteria and alternative is specified as x_ij. Therefore, we have a matrix (x_ij) _ (m*n) expressed as follows:

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

Table 8 Multi-criteria problem example

Ai Cj	*C1*	*C2*	*C3*	*C4*	*C5*	*C6*
A1	*2*	*1500*	*20,000*	*5.5*	*5*	*9*
A2	*2.5*	*2700*	*18,000*	*6.5*	*3*	*5*
A3	*1.8*	*2000*	*21,000*	*4.5*	*7*	*7*
A4	*2.2*	*1800*	*20,000*	*5*	*5*	*5*

where A_1, A_(2),….,A_m are possible alternatives the decision makers want to rank (i.e. classification models). C_1,C_(2),…,C_n are the criteria against which the performance of each alternative is evaluated. Lastly, 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. Certain processes must be achieved to score the alternatives, such as normalisation, maximisation indicator, adding weights and other processes depending on the method. For example, suppose that D is the decision matrix utilised to score the performance of the alternative Ai, where based on C_j, Table 8 is an example of the multi-criteria problem described by [115].

The values in the graph are difficult to evaluate owing to the large numbers in c2 and c3 (Fig. 10).

Enhancing the decision-making process is important by involving the decision makers and stakeholders. Using proper methods for decision making is also necessary to handle multi-criteria problems [210]. MCDM is extensively used in healthcare [112, 116] [116]. Decision makers in healthcare can improve their decision making through a systematic method and attainment of the best decision, depending on the various MCDM methods [116]. In particular, many of the healthcare decisions are complex and unstructured [116].

MCDM Methods

Several MCDM theories have been explored. Figure 11 shows the most commonly used MCDM techniques, which use different notations [116–133]. Table 9 provides a brief description of each technique.

MCDM techniques are diverse, and this variety might cause difficulty in selecting the suitable techniques among the many available MCDM techniques. Each technique has

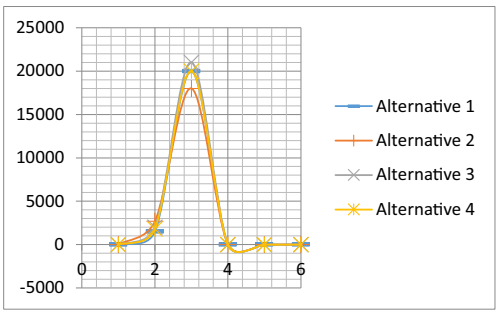

Fig. 10 Graphic illustration of the example in Table 11

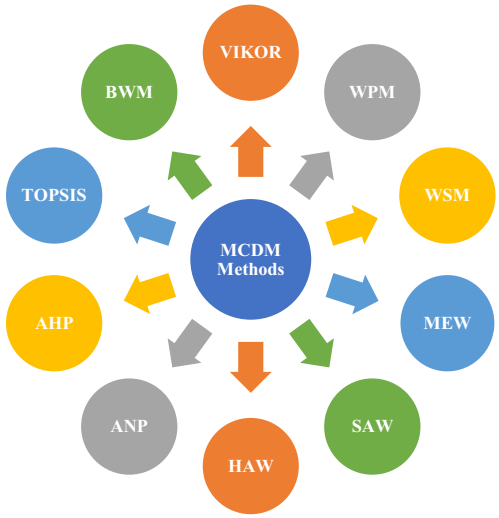

Fig. 11 Commonly used MCDM methods

its own limitations and strengths [118, 134–137]. Thus, selecting the appropriate MCDM method is important. To determine the best MCDM method, many studies have presented their advantages and limitations. Other studies presented a comparison analysis among various MCDM methods. Figure 12 illustrates the advantages and disadvantages of the common MCDM methods [116, 119, 125, 128, 130, 131, 134, 136, 138–168].

To the best of our knowledge, none of the analysed methods have been employed to rank the multiclass classification models for acute leukaemia. Many MCDM methods have been proposed and employed in various studies for the calculation of criteria (factors) weights, such as WSM, SWA, AHP, and BWM [134, 156–159]. The present study has employed the best and worst methods because they can provide the results with more consistency against AHP and other weighting MCDM methods; in addition, the pairwise comparisons based on BWM are lesser than the other methods [141–144]. The pairwise comparisons in BWM method focus on reference comparisons as well; meaning, it executes the preference of the most important criterion over all the other criteria and the preference of all the criteria over the least important criterion [143, 169] [162]. Conversely, the most common MCDM methods for ranking the alternatives are TOPSIS and VIKOR; these two methods employ the compromise priority approach for multiple response optimisation [140] [131] [149]. VIKOR and TOPSIS are both based on an aggregating function representing 'closeness to the ideal'. The ranking index of VIKOR is based on a particular measure of 'closeness' to the ideal solution. Conversely, TOPSIS determines the selected alternative on the basis of its proximity to the ('shortest distance') ideal solution and the greatest distance from the 'negative-ideal' solution; however, it does not consider the relative importance of the distances from these points [131, 138]. In addition, VIKOR can rank the alternatives to determine the best one accurately and rapidly [140]. The recent style of VIKOR studies has changed into integrating VIKOR with another MCDM method rather than applying it alone. In the reviewed studies, numerous examples of applying VIKOR with BWM were provided to achieve consistency improvement for the subjective weights. Such integration between VIKOR with BWM also archives a robust method based in the advantages of the two methods to overcome the uncertainties associated with the problem under study [138, 163–168], VICOR and BWM are clear and easy to use for those without an MCDM background; it can also be performed in a friendly computing environment [138].

Thus, VIKOR and BWM have been adopted to resolve different real-world issues. However, VIKOR cannot elicit the weights and check the decision-making consistency. To overcome these limitations, several authors have recommended employing the BWM with VIKOR [118, 144, 166]; BWM can set weights and check the consistency, along with its flexibility to be applied with other methods. As a conclusion, evaluating and benchmarking acute leukaemia multiclass classification suggest a need to integrate the BWM to set weights for evaluation and benchmarking criteria (reliability, time complexity rate), depending on the judgments of experts. Moreover, VIKOR is recommended to supply the ranking of multiclass classification models. Figure 13 illustrates the proposed solution for the evaluation and benchmarking of acute leukaemia multiclass classification.

Best-worst method (BWM) Determining the most important and desirable alternative is the main aim of the MCDM methods when multiple criteria for decision making exist. Weights are elicited to the decision criteria depending on a comparison among them [143]. Pairwise comparison among the attributes enables us to set the weights for the attributes in each aspect. The BWM method is one of the common multicriteria decision-analysis methods, which perform less pairwise comparison that leads to obtaining the highest consistency in the weight obtaining process [160]. In 2015, Rezaei developed the BWM, in his method amid to weights obtaining for decision criteria and alternatives with respect to multiple various criteria through pairwise comparisons, but it requires a number of comparisons; consequently, it focused on improving the consistency for the weight setting process [169] [155]. The weight elicitation process in BWM depends on reference comparisons, which lead to less comparisons; accordingly, it focuses on determining the best criterion, the preference of this criterion over all the other criteria and the preference of all the criteria over the worst criterion [169]. BWM uses a scale from 1 to 9 to determine the preferences among the criteria. It achieves more reliable outcomes than most MCDM, is easy to use, decreases the times of comparison and ensures the results' reliability by making fewer comparisons [143, 170]. It likewise contains an consistency index

Table 9 Common MCDM Techniques

Methods		brief description
Analytic Hierarchy Process	AHP	AHP reflects the natural behavior of human thinking. It solves complex problems by decomposing them into a hierarchy of more easily comprehended sub-problems having decision alternatives at the lowest levels [110, 111]. It is a popular MCDM method and obtains ratio scales from paired comparisons. It allows small inconsistencies in judgment because humans are precisely consistent [112].
Analytic Network Process	ANP	ANP is defined as a mathematical theory that can handle all types of dependencies systematically. It can be used in numerous fields. ANP includes a multi-criteria decision-making method that compares different alternatives to select the best alternative. ANP technique allows the addition of an extra relevant criterion to an existing one, which are either tangible or intangible, thus significantly influencing the decision-making process [113].
Simple Additive Weighting	SAW	The basic logic of SAW is to obtain the weighted sum of the performance ratings of each alternative over all attributes by performing the following steps [114, 115]. The SAW consists of two basic steps scale the values of all attributes to make them comparable; sum up the values of the all attributes for each alternative [115].
Hierarchical Adaptive Weighting	HAW	In SAW, each criterion value is divided by the largest criterion value among all alternatives. Unlike SAW, the HAW method (20).
Weighted Sum Model	WSM	The WSM is the one of the earliest and probably the simplest technique that is used in MCDM. Due to its simplicity, the technique is suitable for simple problems, as it basically supports single dimensional problems. WSM allows the comparison of the alternatives by assigning scores, and then using these scores, standard values are generated for the alternatives under consideration. The criteria are given weights depending on the severity of each; sum of all these weights must be 1. Each alternative is assessed with respect to every attribute [116, 117]
Weighted Product Method	WPM	It is almost similar to WSM; the only difference between both methods is that addition is the main mathematical operation in WSM, whereas multiplication is the main mathematical operation in WPM [114],Alternatives are being compared with the other by the weights and ratio of one for each criterion [107]
Multiplicative Exponential Weighting	MEW	The main idea of this technique is to take the exponential of each criterion to the weight rather than multiplying the criteria by the weight. Following this step, all the new value of the criterion is aggregated by multiplying the result of the previous step [118].
Vlse Kriterijumska Optimizacija Kompromisno Resenje	VIKOR	The VIKOR method as a typical MCDM method is capable of dealing with the discrete decision-making problems with noncommensurable (different units) and conflicting criteria, and it can help the decision-makers to determine the compromise solution for the problems with multiple conflicting criteria [119, 120].
The technique for order preference by similarity to ideal solution	TOPSIS	TOPSIS is one of the well-known classic MCDM methods. It is a widely accepted multi-attribute decision-making technique due to its sound logic. This technique is based on the concept that the ideal alternative has the best level for all attributes, whereas the negative ideal is the one with all of the worst attribute values [121, 122].
Best-Worst-method	BWM	"BWM is a comparison-oriented MCDM method that compares the best factor to the other factors and all the other factors to the worst factor" [123]

to measure the reliability of the reference comparisons among the criteria. As BWM noted, selecting the best criteria is not difficult among the available criteria, whereas the difficult part is how to determine the importance level of the best criterion over the other criteria, as well as the importance of all the criteria over the worst one. The BWM provides the comparison outcome through the numbers from f 1, 2, 3, 4, 5, 6, 7, 8 and 9 g and neglects the reciprocals of each pair to overcome the problem arising from the unequal distance between fractional comparisons [162].

In general, the BWM focuses on eliciting the weights for decision criteria based on the reference comparison for the most desirable criterion (best one) and the least desirable criterion (worst one) with the other set of decision criteria. Using BMW to elicit the weight includes five sequential steps [138, 143, 160, 166, 171, 172], as demonstrated in Fig. 13.

Table 10 Index of Consistency

aBW	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

VIKOR method VIKOR is one of the common MCDM methods that aim to improve the solutions of complex decisions. VIKOR was developed by Serafim Opricovic under the name 'VlseKriterijumska Optimizacija I Kompromisno Resenje' [88]. Ranking and selecting the alternatives are the

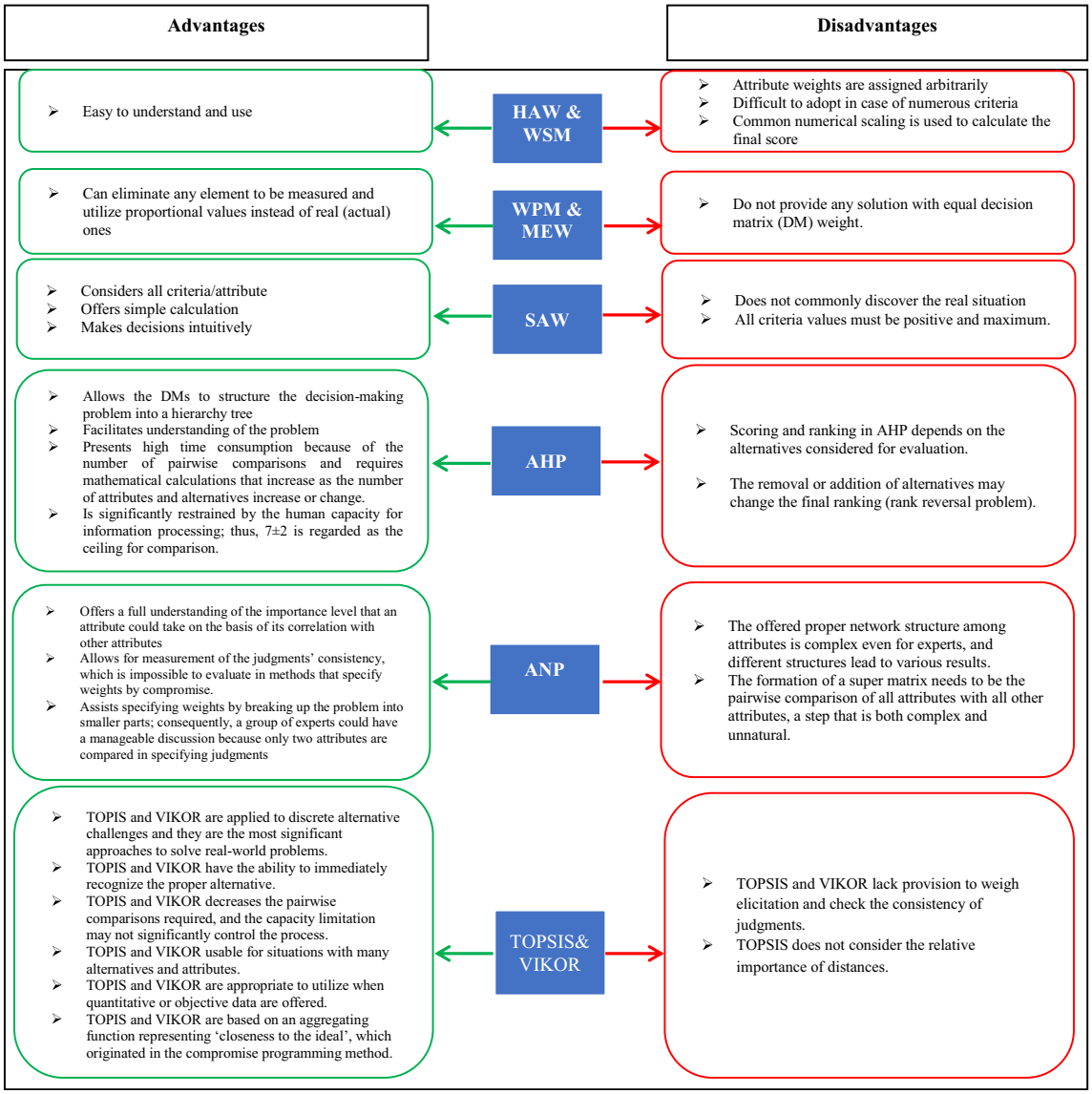

Fig. 12 Advantages and Disadvantages of MCDM Methods

main issues of this method, especially when difficulties arise in decision making because of multiple conflicting criteria [131] [173]. The main principle of VIKOR's work is comparing available alternatives on the basis of the multiple criteria to rank the alternatives and select the best among them. It can deal with criteria even when different measurement units are used [174]. The compromise ranking of alternatives is conducted according to the closeness of the alternative to the ideal solution; meaning, the alternatives will be ranked from the nearest to the ideal solution to the farthest one [175]. VIKOR method must determine the items in compromise ranking, that is, the solution that considers compromising and stabilising the intervals of the weight for preference stability of the compromise solution [131]. For alternative ranking, VIKOR is considered a common method and has been employed by many studies [173, 176, 177].

In VIKOR method, the alternatives and decision criteria are arranged in a structure known as a decision matrix; the columns of this matrix depict a number of alternatives, and the rows represent a number of decision criteria [166, 174].

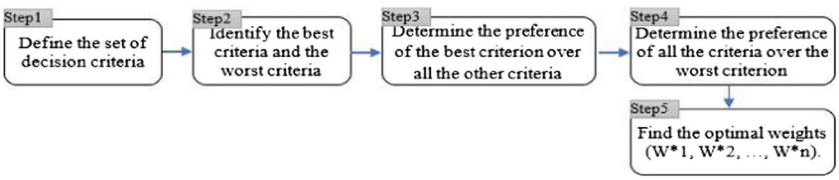

Fig. 13 Steps of the BWM method

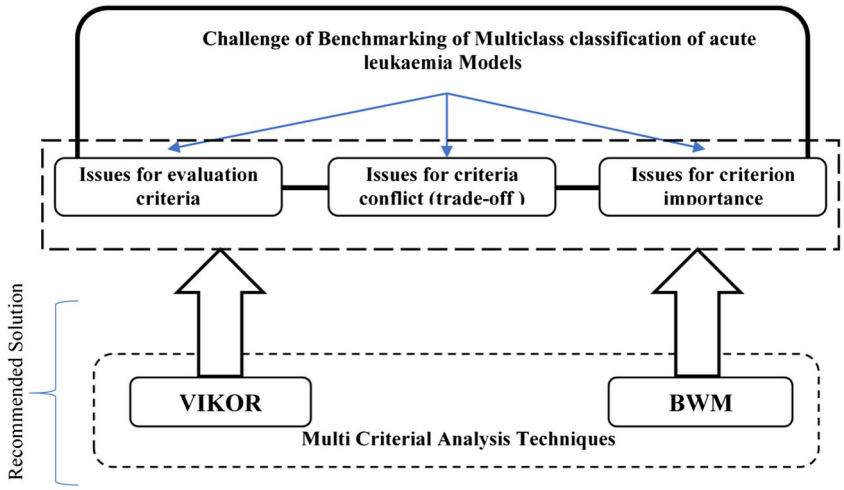

Fig. 14 Proposed solution for the evaluation and benchmarking of acute leukemia classification

The steps in the VIKOR method [88, 131, 175, 176, 178] are as follows:

Step 1. For each criterion, define the highest and lowest values.
Step 2. Construct the weighted decision matrix.
Step 3. Calculate Si and Ri in rough number.
Step 4. Calculate Qi in rough number.
Step 5. Perform alternative ranking.
Step 6. Check the 'acceptable advantage' and 'acceptable stability' in decision making.

Figure 14 illustrates the integrated MCDM method employed as the recommended decision support solution to deal with the main issues of benchmarking/selection of the multiclass classification of acute leukaemia. BWM method is employed to elicit the weights for the evaluation criteria. The weights obtained from BWM are passed to the VIKOR method, which is responsible for the ranking among the alternatives based on weighted criteria.

Methodology aspects

This section introduces the description and explanation of the methodological aspects of the decision support system for evaluating and benchmarking the multiclass classification of acute leukaemia. The identification of the decision matrix based on the evaluation and benchmarking criteria is the first phase (Section "Identification of a decision matrix"), followed by the development phase of a new decision support system for the evaluation and benchmarking based on integrated BWM and VIKOR (Section "Development phase"). The final phase is the validation process (Section "Validation phase"). The proposed methodology is presented in Fig. 15.

Identification of a decision matrix

This phase aims to construct the decision matrix, which is the main component in our decision support system. The decision matrix components consist of decision alternatives and decision criteria. The decision alternatives in our case are multiclass classification models for acute leukaemia, and the criteria are evaluation criteria and sub-criteria identified in the previous phase. Eight multiclass classification models will be built for acute leukaemia to use these models as decision alternatives in our decision matrix.

The identification phase has three steps: The first step builds the multiclass classification models using common types of machine learning methods. The second step performs a cross-over in the decision matrix among various evaluation criteria with different multiclass classification models. The third step evaluates the multiclass classification models on the basis of two groups of criteria. The output of this stage is a proposed decision matrix. The following sections will discuss this phase in further detail.

Building the multiclass classification models

In general, building the multiclass classification model includes a two-step process. The first step is known as the training (learning) process, a model which describes that a predetermined class set is built by analysing the instances of training dataset. Each individual instance is supposed to belong to a predefined class. Each instance is assumed to belong to a predefined class. In the second process, the classification model runs using other independent dataset known as a testing dataset to perform an estimation of the classification model. If the classification model performance looks 'acceptable', the classification model can be used to classify future data for which the class label is unknown. Ultimately, the classification model that provides an acceptable result can be considered an 'acceptable model'.

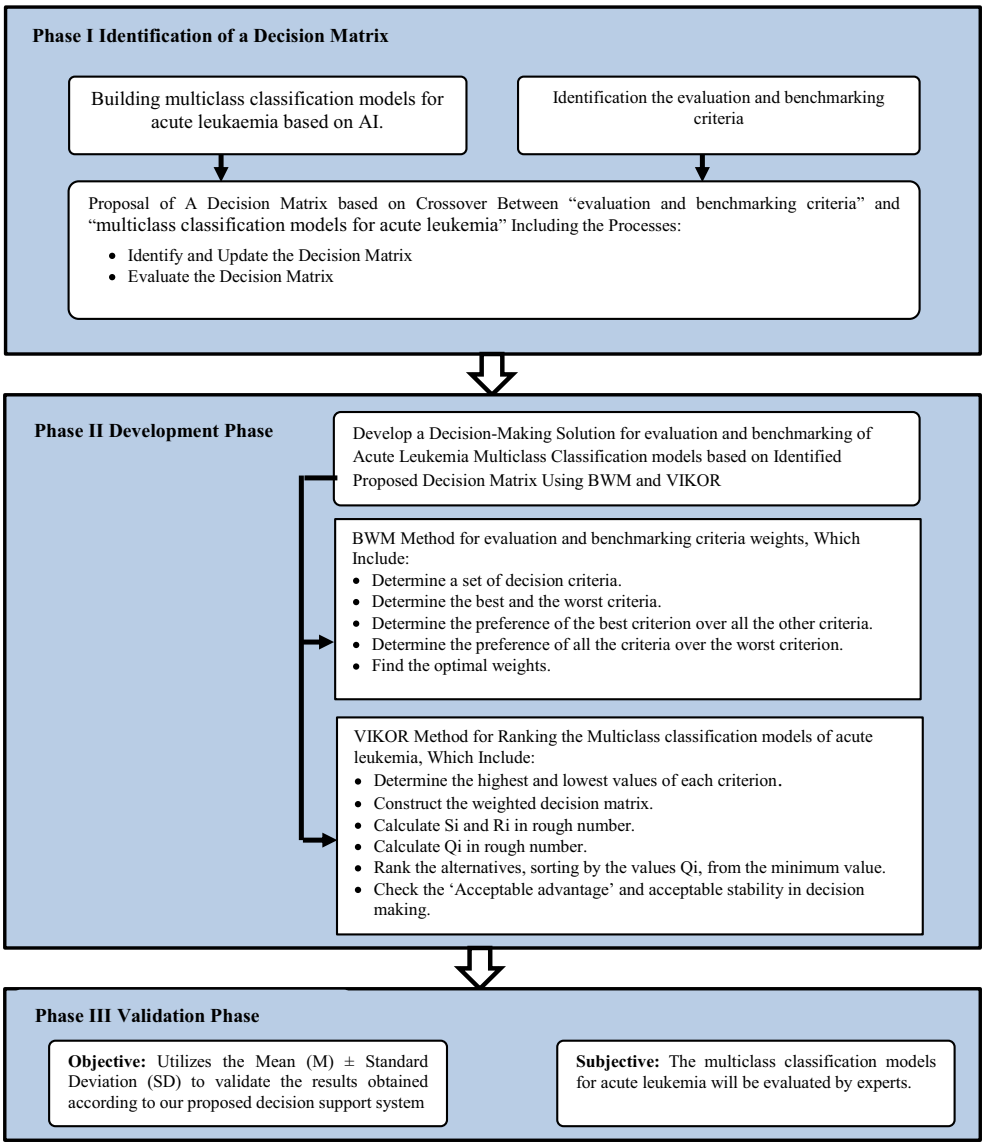

Fig. 15 Methodology of proposed decision support system

The requirements of multiclass classification models and the procedures that will be followed to build multiclass classification models will be described in the following subsections.

Dataset description We adopted the acute leukaemia public microarray dataset proposed by [179]. This dataset is one of the most common datasets in literature, the most frequently used in the papers reviewed and is publicly available. The dataset contains three categories of acute leukaemia, namely, acute myelogenous leukaemia (AML), ALL B cell and ALL T cell. The total number of items in this dataset is 72 samples and 5327 gens.

Multiclass classification processes The multiclass classification processes of microarray datasets of acute leukaemia based on machine learning consist of the following two main processes [18, 74, 103]:

A. Feature Selection. In general, the microarray data have dozens of sample sizes (small) but contain high dimensionality (thousands of genes); however, few parts of the genes affect the result of classification, which means that most genes have no value in classification. Irrelevant genes not only confuse the classification process but also have a negative effect on the classification performance. In addition, over-fitting may occur as a result of irrelevant genes. By contrast, reducing the number of genes has a positive effect on decreasing the input computing; it will also have a positive effect on the overall classification results and performance [23, 69, 81].

In this study, we select a small number of genes that are highly relevant with classification classes, known as informative genes. The chi-square (X^2) method evaluates features individually by measuring their chi-squared statistic with respect to the classes. The X^2 value,

$$x^2(a) = \sum_{v=V}\sum_{i=1}^{n} \frac{[A_i(a=v) - E(a=v)]2}{E_i(a=v)} \quad (9)$$

where V is the set of possible values for a, n is the number of classes, A_i (a = V) is the number of samples in the ith class with $a = v$ and Ei $(a = v)$ is the expected value of A_i $(a = v)$; $Ei(a = v) = P(a = v)$ 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 [20].

The next stage of the classification processes is passing the best subset of features onto a classification model that analyses the quantified characteristics and classifies the data into classes AML, ALL-B_cell and ALL-T_cell.

B. **Multiclass Classification Model** is the process of classifying input patterns to one of a predefined set of classes (e.g. AML, ALL-B_cell and ALL-T_cell) on the basis of the best subset of features, which have been selected in the feature selection stage [74].

In this stage, six multiclass classification models are built based on six well-known machine learning methods that support multiclass classification. These machine learning methods have been extensively employed in previous studies, and all have shown good results when used in the classification microarray dataset; they include ANN, SVM, Decision Tree, Nearest Neighbour, Random Forest and Bayesian Network. The following details all concern each method.

To build multiclass classification models should the dataset be divided into two parts, one part will be used as a training set and the other will be kept as a test set. The training set is used in training the multiclass classification models, and the othe part of the dataset (test set) is used in testing the trained models. Figure 16 illustrates the two processes to bulid the multiclass classification models.

The six built multiclass classification models classify the test dataset into three sets, namely, AML, ALL-B_cell and ALL-T_cell.

Crossover between multiclass classification models and different criteria

The alternatives and criteria are the main components in the decision matrix that will be built. This section describes the procedure of a cross-over between the different alternatives with different criteria. The alternatives are eight classification models built in the previous stage, and the criteria are fifteen of the criteria gathered from the literature review. Figure 17 presents the structure of the decision matrix.

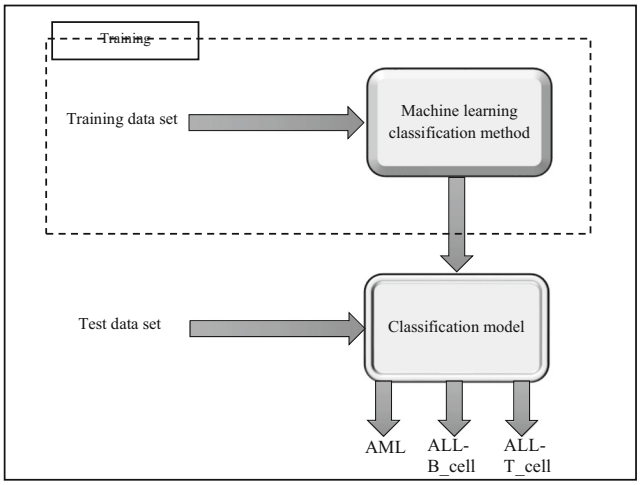

Fig. 16 Multi-class classification model

The alternatives and criteria are the main components in the decision matrix that will be built. This section describes the procedure of a cross-over between the different alternatives with different criteria. The alternatives are eight classification models built in the previous stage, and the criteria are fifteen of the criteria gathered from the literature review. Figure 17 presents the structure of the decision matrix.

Table 16 shows the structure of the proposed decision matrix; the reliability and time complexity are the key sets of criteria with different classification models as alternatives. The next section will discuss the procedures for each criterion in detail.

Computing procedure for reliability group elements The matrix, relationship and behaviour of parameters and error rate represent the four sets of sub-criteria in the reliability group. Firstly, we will generate the first sub-criteria (i.e. confusion matrix) that contain the four main parameters (TP, TN, FN and FP); these parameters represent the basic four criteria in the reliability group. The rest of the sub-criteria within the reliability group are calculated based on the confusion matrix parameters by a certain formula.

Thus, the values of each multiclass classification model will be calculated separately by conducting experiments to generate final parameter values for the decision matrix.

Computing procedure for the time complexity criterion The calculating procedure for the time complexity is based on the time consumed by the input of the sample dataset and the output of the results (Fig. 18).

The procedure of calculating the sample process depends on the number and size of samples through the following question:

$$T_{process} = T_o - T_i, \quad (10)$$

where T_o is the output time image process, and T_i is the input time sample process:

Criteria

Alternatives: Classification Models	Reliability				Time Complexity
Model 1 (ANN)	RV (M1/ TS)	MPV (M1/ TS)	BPV (M1/ TS)	ERV (M1/ TS)	TcV (M1/ TS)
Model 2 (NN)	RV (M2/ TS)	MPV (M2/ TS)	BPV (M2/ TS)	ERV (M2/ TS)	TcV (M2/ TS)
Model 3 (SVM)	RV (M3/ TS)	MPV (M3/ TS)	BPV (M3/ TS)	ERV (M3/ TS)	TcV (M3/ TS)
Model 4 (RF)	RV (M4/ TS)	MPV (M4/ TS)	BPV (M4/ TS)	ERV (M4/ TS)	TcV (M4/ TS)
Model 5 (DT)	RV (M5/ TS)	MPV (M5/ TS)	BPV (M5/ TS)	ERV (M5/ TS)	TcV (M5/ TS)
Model 6 (BN)	RV (M6/ TS)	MPV (M6/ TS)	BPV (M6/ TS)	ERV (M6/ TS)	TcV (M6/ TS)
…	…	…	…	…	…
…	…	…	…	…	…
Model n	RV (Mn/ TS)	MPV (Mn/ TS)	BPV (Mn/ TS)	ER (Mn/ TS)	TcV (Mn/ TS)

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

Fig. 17 Structure of decision matrix

$$T_{total} = \frac{T_{process}}{T_{Avarge}} \tag{11}$$

where $T_{process}$ is the difference among output and input samples, and $T_{average}$ represents the average sample process for all samples.

Evaluation step

Multiple criteria and different multiclass classification models are the main components of the decision matrix; multiple criteria represent the two main criteria (i.e. reliability and time complexity) used to evaluate the multiclass classification models. The alternatives in the decision matrix are eight multiclass classification models. The decision matrix tested and evaluated based on the calculation procedure for classification models is performed from the experiments for data collection, which expresses the decision matrix final results.

Development phase

This phase focuses on achieving the third objective of the research, that is, 'to develop a new decision support system for evaluation and benchmarking of acute leukaemia multiclass classification using multi-criteria decision-making techniques'. Thus, a new decision support system will be developed based on MCDM techniques. The integration of BWM and VIKOR techniques will be adopted for ranking and selecting the best alternatives on the basis of the decision matrix built in the previous phase. The steps of this phase will be presented below. Figure 19 illustrates the new decision support system for the evaluation and benchmarking of multiclass classification model of acute leukaemia.

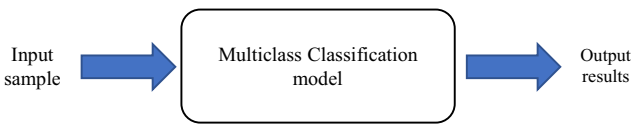

Fig. 18 Computing the time complexity

Figure 18 demonstrates the components of new decision support system for the evaluation and benchmarking of multiclass classification models for acute leukaemia. The alternatives are ranked to determine which of them is the best using the integrated methods of BWM and VIKOR.

Developing integrated methods of BWM and VIKOR for the evaluation and benchmarking/selection using MCDM

The MCDM literature suggests that the newest trend in MCDM studies is integrating two or more methods. This integrated approach takes the advantages of two methods and overrides the limitations of using one method. The BWM and VIKOR methods are most extensively used owing to their many advantages. Combining BWM and VIKOR methods is also widely accepted in literature because of their capacity to present the results of complete ranking and calculate the relative distance on the basis of weights and objective data. A new decision support system for evaluation and benchmarking is designed based on the integration of BWM and VIKOR techniques (Fig. 18). The alternatives include the multiclass classification models. The criteria are the reliability group, time complexity and weight (human preferences). Those components represent the decision matrix.

BWM In this stage, several steps are involved to assign proper weights to the multi-service criteria using BWM. The BWM procedure includes the following steps [143, 180].

Step 1. Determine a set of decision criteria

The first step of the BWM is determining the criteria set, C1, C2, …; Cn should be used by the decision maker when deciding on the best alternative. In our study, the criteria set is

Criteria

Alternatives: Classification Models	Reliability				Time Complexity
Model 1 (ANN)	RV (M1/ TS)	MPV (M1/ TS)	BPV (M1/ TS)	ERV (M1/ TS)	TcV (M1/ TS)
Model 2 (NN)	RV (M2/ TS)	MPV (M2/ TS)	BPV (M2/ TS)	ERV (M2/ TS)	TcV (M2/ TS)
Model 3 (SVM)	RV (M3/ TS)	MPV (M3/ TS)	BPV (M3/ TS)	ERV (M3/ TS)	TcV (M3/ TS)
Model 4 (RF)	RV (M4/ TS)	MPV (M4/ TS)	BPV (M4/ TS)	ERV (M4/ TS)	TcV (M4/ TS)
Model 5 (DT)	RV (M5/ TS)	MPV (M5/ TS)	BPV (M5/ TS)	ERV (M5/ TS)	TcV (M5/ TS)
Model 6 (BN)	RV (M6/ TS)	MPV (M6/ TS)	BPV (M6/ TS)	ERV (M6/ TS)	TcV (M6/ TS)
…	…	…	…	…	…
…	…	…	…	…	…
Model n	RV (Mn/ TS)	MPV (Mn/ TS)	BPV (Mn/ TS)	ER (Mn/ TS)	TcV (Mn/ TS)

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

Fig. 19 New decision support system for the evaluation and benchmarking of the multiclass classification models for acute leukaemia

obtained from the analysis conducted in the literature. Figure 20 illustrates the main and sub-criteria, which are dependent in this study.

Step 2. Determine the best and worst criteria

The best criterion can be considered as the most desirable or the most important criteria to the decision, and the worst criterion represents the less desirable or less important criteria to the decision. In this step, the best and the worst criteria are defined depending on the decision maker/evaluator's perspective.

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

The process of pairwise comparison is conducted between the identified best criterion and the other criteria. This step aims to determine the preference of the best criterion over all the other criteria. The evaluator/expert must determine a value from 1 to 9 to represent the importance of the best criterion over the other criteria.

This procedure will result in a vector known as 'Best-to-Others', which is

$$AB = (a_{B1}, a_{B2}, \ldots a_{Bn},) \quad (12)$$

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

Step 4. Process the pairwise comparison between the other criteria and the worst criterion

This comparison aims to identify the preference of all criteria over the least important criterion. The evaluator/expert determines the importance of all the criteria over the worst criterion; the numbers from 1 to 9 are used to indicate the importance. The result of this step is a vector known as 'Others-to-Worst'.

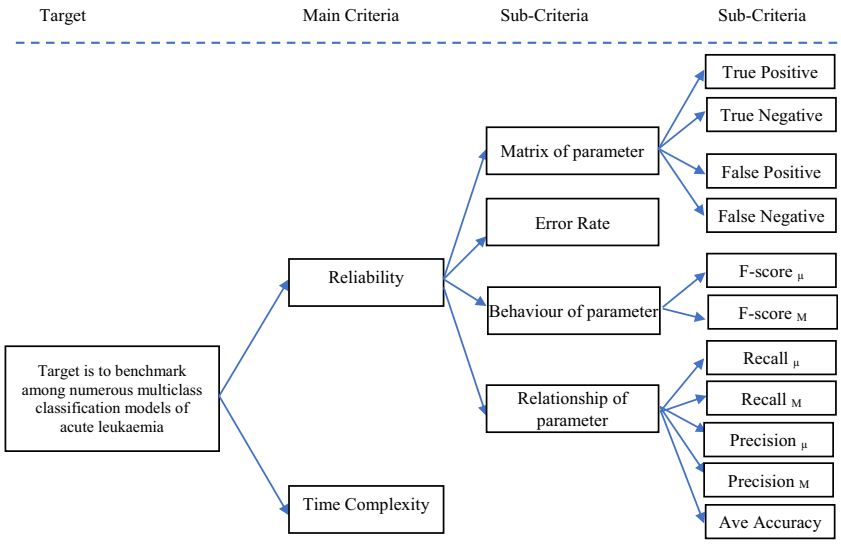

Fig. 20 Main and sub-criteria used

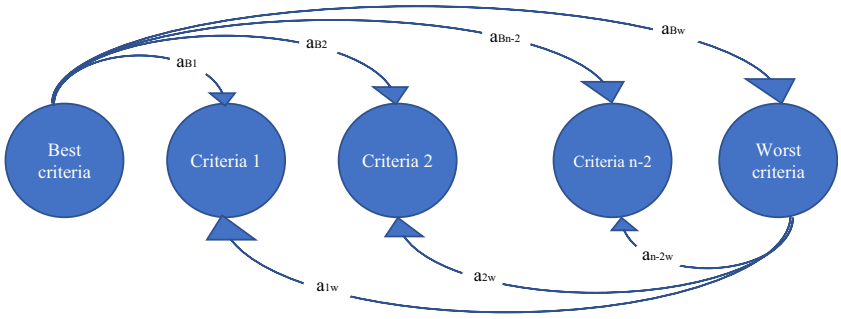

Fig. 21 Reference Comparisons in the BWM Method

Others-to-Worst vector result 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, $a_{ww} = 1$.

Figure 21 illustrates the two types of reference comparisons, namely, Best-to-Others and Others-to-Worst criteria.

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

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

To fulfil these conditions for all j, we should find a solution where the maximum absolute differences

$$\left|\frac{W_B}{W_j} - a_{Bj}\right| and \left|\frac{W_i}{W_w} - a_{Bjw}\right| \quad (13)$$

for all j are minimised. Considering the non-negativity and sum condition for the weights, the following problem is created:

$$\begin{aligned} &\min\max_j \left\{ \left|\frac{W_B}{W_j} - a_{Bj}\right|, \left|\frac{W_j}{W_w} - a_{jw}\right| \right\} \\ \text{s.t.} \quad &\sum_j W_{j=1} \\ &W_j \geq 0, \text{for all } j \end{aligned} \quad (14)$$

Problem (5) can be transferred to the following problem:

$\min \xi$,

s.t.

$$\left|\frac{W_B}{W_J} - a_{Bj}\right| \leq \xi, for\ all\ j \quad (15)$$

$$\left|\frac{W_j}{W_w} - a_{jw}\right| \leq \xi, for\ all\ j \quad (16)$$

$\sum_j W_{j=1}$

By solving Problem (6), the optimal weights ($w^*_1; w^*_2;\ldots;w^*_n$) 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 a high consistency and thus a high reliability.

Then, the consistency ratio is calculated using ξ* and the corresponding consistency index as follows:

$$\text{Consistency Ratio} = \frac{\xi^*}{Consistency\ Index} \quad (17).$$

As proposed by [143], clearly, the bigger the ξ*, the more consistent the vectors are.

Adaptive VIKOR method for multiclass classification model ranking To rank multiclass classification models in this phase, VIKOR is utilised because of its suitability for the decision case with many alternatives and multiple conflicting criteria. In addition, it is capable of providing results rapidly while determining the most appropriate option. All the criteria weights will also be obtained from the WBM and will be used in the VIKOR. The available decision alternative results are ranked in decreasing order, and the hospitals are ranked on the basis of the number of available services employing the VIKOR method. Based on the VIKOR method, the multiclass classification models are ranked according to the identified weighted criteria.

VIKOR steps are presented in the following [242], [273].

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

$$f_i^* = \max_j f_{ij}, f_i^- = \min_j f_{ij}. \quad (18)$$

Step 2: The weights for each criterion are computed based on the BWM method. 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.

$$WM = wi*\left(f^*i-fij\right)/\left(f^*i-f^-i\right) \tag{19}$$

This process will produce a weighted matrix as follows:

$$\begin{bmatrix} w_1\left(f^*1-f11\right)/\left(f^*1-f^-1\right) & w_2\left(f^*2-f12\right)/\left(f^*2-f^-2\right) & \dots & w_i\left(f^*i-fij\right)/\left(f^*i-f^-i\right) \\ w_1\left(f^*1-f21\right)/\left(f^*1-f^-1\right) & w_2\left(f^*2-f22\right)/\left(f^*2-f^-2\right) & \dots & w_i\left(f^*i-fij\right)/\left(f^*i-f^-i\right) \\ \vdots & \vdots & \vdots & \\ w_1\left(f^*1-f31\right)/\left(f^*1-f^-1\right) & w_2\left(f^*2-f32\right)/\left(f^*2-f^-2\right) & \dots & w_i\left(f^*i-fij\right)/\left(f^*i-f^-i\right) \end{bmatrix} \tag{20}$$

Step 3: Compute the *Sj* and *Rj* values, $j = 1,2,3,\ldots,J$, $i = 1,2,3,\ldots,n$ by using the following equations:

$$Sj = \sum_{i=1}^{n} wi*\left(f^*i-fij\right)/\left(f^*i-f^-i\right) \tag{21}$$

$$Rj = \max_i wi*\left(f^*i-fij\right)/\left(f^*i-f^-i\right) \tag{22}$$

where *wi* are the weights of criteria expressing their relative importance.

Step 4: Compute the values Q_j, $j = (1, 2, \cdots, J)$ by the following relation:

$$Q_j = \frac{v\left(S_j-S^*\right)}{S^--S^*} + \frac{(1-v)\left(R_j-R^*\right)}{R^--R^*} \tag{23}$$

where

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

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

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

Step 5: The set of alternatives (hospitals) can now be ranked by sorting the values S, R and Q in ascending order. The lowest value indicates the optimal performance.

Step 6: Propose as a compromise solution the alternative (a'), which is ranked the best by the measure Q (minimum) if the following two conditions are satisfied:

- C1. 'Acceptable advantage':

$$\left(Qa''\right)-\left(Qa'\right) \geq DQ,$$

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

- 'Stability' is acceptable with the decision-making context: Alternative a' should also be the best ranked by S and/or R.

This compromise solution is stable within the process of decision making, which could 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').

Validation phase

Selection decisions for the multiclass classification model is a difficult task because they depend on conflicting multiple criteria in one side and the difference among them regarding performance and accuracy and other features. The proposed decision support results are validated by utilising subjective and objective validations.

Objective validation

To ensure the ranking of multiclass classification models on the basis of the proposed decision support system, this study employs two statistical methods (mean ± standard deviation). The ranking results of multiclass classification models are separated into four similar groups [181, 182]. The results of each group are expressed as mean ± standard deviation.

Mean is the result average. It is calculated by dividing the sum of the observed results over the number of results and by the following equation:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{24}$$

Standard deviation is utilised to determine the dispersion or variation amount in the set of values and is calculated by the following equation:

$$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(x_i-\overline{x}\right)^2} \tag{25}$$

Mean ± SD is utilised to ensure that the four sets of multiclass classification models are subject to systematic ordering. To validate the ranking results using the above test, the scoring of the multiclass classification models was split into

four groups on the basis of the ranking result obtained from the proposed decision support system. Each group contains an equal number of selected multiclass classification models (2) depending on the scoring values from the ranking results. This process is performed using two methods based on a statistical platform, which must prove that the first group reached the highest scoring value when the mean and SD were measured. For result validation, we assumed that the first group had the highest mean and SD than the other three groups. The second group's mean and SD results must be lower than or equal to those of the first group. However, the third group's mean and SD results must be lower than those of the first and second groups or equal to those of the second group. Lastly, the fourth group's mean and SD must be lower than those of first, second and third groups or equal to those of the third group. According to the systematic ranking results, the first group must be proven statistically to be considered the highest group among other groups.

Subjective validation

This section describes the subjective validation process. The multiclass classification models will be evaluated by the specialist experts in the data classification of medical cases by machine learning. The experts prove the effectiveness of the multiclass classification models' ranking obtained by our proposed decision support system. They decide the validated ranking of multiclass classification models on the basis of our decision support system by examining the values of all the evaluation criteria used.

Conclusion

Studies on automated detection and classification of acute leukaemia have increased. However, studies relevant to the evaluation and benchmarking of automated detection and classification tasks whose limitations have remained unaddressed are scarce. The evaluation and benchmarking process of automated detection and classification involve numerous aspects that require further investigation and in-depth analysis. A comprehensive review of relevant literature is the key contribution of this study. A systematic search and review were performed for the automated detection and classification of acute leukaemia, focusing on its evaluation and benchmarking aspects, to determine the research gaps and open challenges and issues in the evaluation and benchmarking process. In the taxonomy analysis results, three groups of articles are organised that represent the main research directions in the automated detection and classification research area, namely, proposed methods, proposals for system development and evaluation and comparative analysis. In addition, we designed a corresponding taxonomy that includes the main criteria for evaluating and benchmarking the detection and classification tasks. The two taxonomies were then linked according to the criteria used in each research direction to illustrate the intensity of use of certain evaluation criteria at the expense of others across the identified research directions, as well as the mapping of the use of criteria to evaluate and benchmark the tasks of detection and classification. We analysed in depth all the evaluation and benchmarking aspects in all reviewed studies to highlight the open challenges and issues related to evaluation and benchmarking and the settling of the research gap. A serious gap was observed in the reviewed studies, which failed to perform the evaluation and benchmarking process of all the major requirements of the detection and classification. They only partially conducted the evaluation and benchmarking, thereby rendering the results incomplete as they failed to reflect the overall performance of detection and classification. Such shortcoming causes a challenge when comparing numerous models or systems of detection and classification to determine which of these systems is the best because the evaluation criteria vary and are incomplete. To fill the research gap and address the challenges and open issues, a decision support system was proposed to evaluate and benchmark the multiclass classification models of acute leukaemia. The methodological aspects were described based on three key phases. The identification of the decision matrix based on evaluation and benchmarking criteria is the first phase, followed by the development phase of a new decision support system for evaluation and benchmarking based on integrated BWM and VIKOR. The final phase is the validation process. The proposed decision support system of evaluation and benchmarking of the automated multiclass classification will be implemented to provide evaluation and benchmarking services during the identified challenges and open issues.

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 institutional 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.

References

1. De Paz, J. F. et al., Biomedic Organizations: An intelligent dynamic architecture for KDD. *Inf. Sci.* 224:49–61, 2013.
2. Chen, X. and Jian, C. A tumor classification model using least square regression. In: *2014 10th International Conference on Natural Computation (ICNC)*. 2014.
3. 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.
4. 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.
5. Fann, Y. C., Enhancing patient care and outcomes through innovative informatics systems and tools. *Comput. Methods Prog. Biomed.* 158:A1, 2018.
6. Torkaman, A., et al. A recommender system for detection of leukemia based on cooperative game. In: *2009 17th Mediterranean Conference on Control and Automation*. 2009.
7. Zhiyong, Y., Jingcheng, L., and Zhang, T., Extreme Large Margin Distribution Machine and its applications for biomedical datasets. In: *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2016.
8. 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.
9. Escalante, H. J. et al., Acute leukemia classification by ensemble particle swarm model selection. *Artif. Intell. Med.* 55(3):163–175, 2012.
10. Labati, R.D., et al., ALL-IDB: The acute lymphoblastic leukemia image database for image processing. In: *2011 18th Ieee International Conference on Image Processing*. 2011.
11. Singhal, V. and Singh, P. Local Binary Pattern for automatic detection of Acute Lymphoblastic Leukemia. In: *2014 Twentieth National Conference on Communications (NCC)*. 2014.
12. 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(6):536–542, 2008.
13. 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.
14. Laosai, J. and Chamnongthai, K., Acute leukemia classification by using SVM and K-Means clustering. In: *2014 International Electrical Engineering Congress (iEECON)*. 2014.
15. 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.
16. Agaian, S., Madhukar, M., and Chronopoulos, A. T., Automated Screening System for Acute Myelogenous Leukemia Detection in Blood Microscopic Images. *IEEE Syst. J.* 8(3):995–1004, 2014.
17. Mohapatra, S., et al., Fuzzy Based Blood Image Segmentation for Automated Leukemia Detection>. In: *2011 International Conference on Devices and Communications (ICDeCom)*. 2011.
18. 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.
19. Srisukkham, W. et al., Intelligent leukaemia diagnosis with bare-bones PSO based feature optimization. *Appl. Soft Comput.* 56: 405–419, 2017.
20. Snousy, M. B. A. et al., Suite of decision tree-based classification algorithms on cancer gene expression data. *Egypt. Informatics J.* 12(2):73–82, 2011.
21. Mishra, S. et al., Gray level co-occurrence matrix and random forest based acute lymphoblastic leukemia detection. *Biomed. Signal Process. Control.* 33:272–280, 2017.
22. Nguyen, T., and Nahavandi, S., Modified AHP for Gene Selection and Cancer Classification Using Type-2 Fuzzy Logic. *IEEE Trans. Fuzzy Syst.* 24(2):273–287, 2016.
23. Lei, X., and Chen, Y., Multiclass Classification of Microarray Data Samples with Flexible Neural Tree. In: *2012 Spring Congress on Engineering and Technology*. 2012.
24. Soares, C., et al., Automating Microarray Classification Using General Regression Neural Networks. In: *2008 Seventh International Conference on Machine Learning and Applications*. 2008.
25. Wang, H.-Q. et al., A neural network-based biomarker association information extraction approach for cancer classification. *J. Biomed. Inform.* 42(4):654–666, 2009.
26. Wang, X., and Wang, S., Enhanced algorithm for high-dimensional data classification. *Appl. Soft Comput.* 40:1–9, 2016.
27. Huang, H.L., et al., Boosting Evolutionary Support Vector Machine for Designing Tumor Classifiers from Microarray Data. In: *2007 IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*. 2007.
28. Kumar, M., and Kumar Rath, S., Classification of microarray using MapReduce based proximal support vector machine classifier. *Knowl.-Based Syst.* 89:584–602, 2015.
29. Nasir, A. S. A., Mashor, M. Y., and Hassan, R., Leukaemia screening based on fuzzy ARTMAP and simplified fuzzy ARTMAP neural networks. In: *2012 IEEE-EMBS Conference on Biomedical Engineering and Sciences*. 2012.
30. Qasem, M., and Nour, M., Improving Accuracy for Classifying Selected Medical Datasets with Weighted Nearest Neighbors and Fuzzy Nearest Neighbors Algorithms. In: *2015 International Conference on Cloud Computing (ICCC)*. 2015.
31. 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.
32. Supardi, N. Z., et al., Classification of blasts in acute leukemia blood samples using k-nearest neighbour. In: *2012 IEEE 8th International Colloquium on Signal Processing and its Applications*. 2012.
33. Campos, L. M. d., et al., Bayesian networks classifiers for gene-expression data. In: *2011 11th International Conference on Intelligent Systems Design and Applications*. 2011.
34. 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.
35. Wang, S.-L. et al., Tumor classification by combining PNN classifier ensemble with neighborhood rough set based gene reduction. *Comput. Biol. Med.* 40(2):179–189, 2010.
36. Fan, Y., Hua-Zhen, W., and Hong, M., A novel classification method of microarray with reliability and confidence. In: *2008 International Conference on Machine Learning and Cybernetics*. 2008.
37. El-Nasser, A. A., Shaheen, M., and El-Deeb, H., Enhanced leukemia cancer classifier algorithm. In: *2014 Science and Information Conference*. 2014.
38. Ren, C. -X., Dai, D. -Q., and Yan, H., Robust classification using L 2,1 norm based regression model. *Pattern Recogn.* 45(7):2708–2718, 2012.
39. Kim, S., Spectral Methods for Cancer Classification Using Microarray Data. In: *2009 International Joint Conference on Computational Sciences and Optimization*. 2009.
40. 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.
41. Lu, H. et al., A cost-sensitive rotation forest algorithm for gene expression data classification. *Neurocomputing.* 228:270–276, 2017.

42. Zhang, L. et al., Similarity-balanced discriminant neighbor embedding and its application to cancer classification based on gene expression data. *Comput. Biol. Med.* 64:236–245, 2015.
43. Chandra, B., and Gupta, M., Robust approach for estimating probabilities in Naïve–Bayes Classifier for gene expression data. *Expert Syst. Appl.* 38(3):1293–1298, 2011.
44. Saengsiri, P., et al., Classification models based-on incremental learning algorithm and feature selection on gene expression data. In: *The 8th Electrical Engineering/ Electronics, Computer, Telecommunications and Information Technology (ECTI) Association of Thailand - Conference 2011*. 2011.
45. Wang, H. -Q. et al., Extracting gene regulation information for cancer classification. *Pattern Recogn.* 40(12):3379–3392, 2007.
46. Rajwa, B. et al., Automated Assessment of Disease Progression in Acute Myeloid Leukemia by Probabilistic Analysis of Flow Cytometry Data. *IEEE Trans. Biomed. Eng.* 64(5):1089–1098, 2017.
47. He, Y., and Hui, S. C., Exploring ant-based algorithms for gene expression data analysis. *Artif. Intell. Med.* 47(2):105–119, 2009.
48. Li, J. -T., and Jia, Y. -M., An Improved Elastic Net for Cancer Classification and Gene Selection. *Acta Automat. Sin.* 36(7):976–981, 2010.
49. Krappe, S., et al., Automated morphological analysis of bone marrow cells in microscopic images for diagnosis of leukemia: Nucleus-plasma separation and cell classification using a hierarchical tree model of hematopoesis. In: Tourassi, G. D., and Armato, S. G., (Eds.), *Medical Imaging 2016: Computer-Aided Diagnosis*. 2015.
50. 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 Evol. Comput.* 28:144–160, 2016.
51. Chakraborty, S., Simultaneous cancer classification and gene selection with Bayesian nearest neighbor method: An integrated approach. *Comput. Stat. Data Anal.* 53(4):1462–1474, 2009.
52. Yongqiang, D., 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.
53. Tran, V. N., et al., An automated method for the nuclei and cytoplasm of Acute Myeloid Leukemia detection in blood smear images. In: *2016 World Automation Congress (WAC)*. 2016.
54. Cui, Y. et al., Sparse maximum margin discriminant analysis for feature extraction and gene selection on gene expression data. *Comput. Biol. Med.* 43(7):933–941, 2013.
55. 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.
56. Hasan, A., and Akhtaruzzaman, A. Md., High dimensional microarray data classification using correlation based feature selection. In: *2012 International Conference on Biomedical Engineering (ICoBE)*. 2012.
57. Cao, J. et al., A fast gene selection method for multi-cancer classification using multiple support vector data description. *J. Biomed. Inform.* 53:381–389, 2015.
58. 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.
59. Paul, S., and Maji, P., Rough set based gene selection algorithm for microarray sample classification. In: *2010 International Conference on Methods and Models in Computer Science (ICM2CS-2010)*. 2010.
60. Mohapatra, P., and Chakravarty, S., Modified PSO based feature selection for Microarray data classification. In: *2015 IEEE Power, Communication and Information Technology Conference (PCITC)*. 2015.
61. Shi, T. W., et al., Random Forest and Gene Ontology for functional analysis of microarray data. In: *2014 IEEE 7th International Workshop on Computational Intelligence and Applications (IWCIA)*. 2014.
62. 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.
63. 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.
64. Lu, X., et al., A novel feature selection method based on CFS in cancer recognition. In: *2012 IEEE 6th International Conference on Systems Biology (ISB)*. 2012.
65. Rosa, J. L. D., et al., 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.
66. Chiang, J. H., and Ho, S. H., A Combination of Rough-Based Feature Selection and RBF Neural Network for Classification Using Gene Expression Data. *IEEE Trans. NanoBiosci.* 7(1):91–99, 2008.
67. Pisharady, P. K., Vadakkepat, P., and Poh, L. A., Fuzzy-rough discriminative feature selection and classification algorithm, with application to microarray and image datasets. *Appl. Soft Comput.* 11(4):3429–3440, 2011.
68. Qizhong, Z., Gene selection and classification using non-linear kernel support vector machines based on gene expression data. In: *2007 IEEE/ICME International Conference on Complex Medical Engineering*. 2007.
69. 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.
70. Begum, S., Chakraborty, D., and Sarkar, R., Data Classification Using Feature Selection and kNN Machine Learning Approach. In: *2015 International Conference on Computational Intelligence and Communication Networks (CICN)*. 2015.
71. Chen, T. C., et al., Feature selection and classification by using grid computing based evolutionary approach for the microarray data. In: *2010 3rd International Conference on Computer Science and Information Technology*. 2010.
72. Roy, A., Mackin, P. D., and Mukhopadhyay, S., Methods for pattern selection, class-specific feature selection and classification for automated learning. *Neural Netw.* 41:113–129, 2013.
73. Maulik, U., Mukhopadhyay, A., and Chakraborty, D., Gene-Expression-Based Cancer Subtypes Prediction Through Feature Selection and Transductive SVM. *IEEE Trans. Biomed. Eng.* 60(4):1111–1117, 2013.
74. 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.
75. Nazlibilek, S. et al., Automatic segmentation, counting, size determination and classification of white blood cells. *Measurement.* 55:58–65, 2014.
76. Rawat, J. et al., Computer Aided Diagnostic System for Detection of Leukemia Using Microscopic Images. *Procedia Computer Science.* 70:748–756, 2015.
77. 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.

78. Dehghan Khalilabad, N., and Hassanpour, H., Employing image processing techniques for cancer detection using microarray images. *Comput. Biol. Med.* 81:139–147, 2017.
79. Putzu, L., Caocci, G., and Di Ruberto, C., Leucocyte classification for leukaemia detection using image processing techniques. *Artif. Intell. Med.* 62(3):179–191, 2014.
80. Shankar, V., et al., Automatic detection of acute lymphoblasitc leukemia using image processing. In: *2016 IEEE International Conference on Advances in Computer Applications (ICACA)*. 2016.
81. Horng, J. -T., et al., An expert system to classify microarray gene expression data using gene selection by decision tree. *Expert Syst. Appl.* 36(5):9072–9081, 2009.
82. 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.
83. Ganesh Kumar, P., et al., Design of fuzzy expert system for microarray data classification using a novel Genetic Swarm Algorithm. *Expert Syst. Appl.* 39(2):1811–1821, 2012.
84. Krappe, S., et al., 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., and Tourassi, G. D., (Eds.), *Medical Imaging 2015: Computer-Aided Diagnosis*. 2015.
85. 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.
86. Hossin, M., and Sulaiman, M., A review on evaluation metrics for data classification evaluations. *Int. J. Data Min. Knowl. Manag. Process.* 5(2):1, 2015.
87. Tai, W.L., et al., Blood Cell Image Classification Based on Hierarchical SVM. In: *2011 IEEE International Symposium on Multimedia*. 2011.
88. Gul, M., et al., A state of the art literature review of VIKOR and its fuzzy extensions on applications. *Appl. Soft Comput.* 46:60–89, 2016.
89. Sokolova, M., and Lapalme, G., A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.* 45(4):427–437, 2009.
90. Saritha, M., et al., 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.
91. Zhang, C., et al., An imbalanced data classification algorithm of improved autoencoder neural network. In: *2016 Eighth International Conference on Advanced Computational Intelligence (ICACI)*. 2016.
92. Zupan, B., and Demsar, J., Open-source tools for data mining. *Clin. Lab. Med.* 28(1):37–54, 2008.
93. Daqqa, K. A. S. A., Maghari, A. Y. A., and Sarraj, W. F. M. A., Prediction and diagnosis of leukemia using classification algorithms. In: *2017 8th International Conference on Information Technology (ICIT)*, 2017.
94. Dwivedi, S., Kasliwal P., and Soni, S., Comprehensive study of data analytics tools (RapidMiner, Weka, R tool, Knime). In: *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*. 2016.
95. Sharma, R., Singh, S. N., and Khatri, S., Medical Data Mining Using Different Classification and Clustering Techniques: A Critical Survey. In: *2016 Second International Conference on Computational Intelligence & Communication Technology (CICT)*. 2016.
96. Cadenas, J. M., Garrido, M. C., and Martínez, R., A tool to manage low quality datasets. In: *2012 IEEE International Conference on Fuzzy Systems*, 2012.
97. Wahbeh, A.H., et al., A comparison study between data mining tools over some classification methods. *Int. J. Adv. Comput. Sci. Appl.* 2011. **Special Issue on Artificial Intelligence: p. 18–26.**
98. Rangra, K., and Bansal, D. K. L., Comparative Study of Data Mining Tools, International Journal of Advanced Research in Computer Science and Software Engineering. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* 4(6), 2014.
99. 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.
100. Rawat, J., et al., Review of leukocyte classification techniques for microscopic blood images. In: *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*. 2015.
101. Bagasjvara, R. G., et al., 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.
102. Singh, G., Bathla, G., and Kaur, S., A review to detect leukemia cancer in medical images. In: *2016 International Conference on Computing, Communication and Automation (ICCCA)*. 2016.
103. Wahhab, H. T. A., *Classification of acute leukemia using image processing and machine learning techniques*. PhD thesis, University of Malaya, 2015.
104. Shafique, S., and Tehsin, S., Computer-Aided Diagnosis of Acute Lymphoblastic Leukaemia. *Comput. Math. Methods Med.* 2018.
105. 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.* IEEE, 2009.
106. Keeney, R. L., and Raiffa, H., *Decisions with multiple objectives: Preferences and value trade-offs*. Cambridge University Press, 1993
107. Belton, V., and Stewart, T., *Multiple criteria decision analysis: An integrated approach*. Kluwer Academic Publishers: Boston, 2002.
108. Petrovic-Lazarevic, S., and Abraham, A., *Hybrid fuzzy-linear programming approach for multi criteria decision making problems.* arXiv preprint cs/0405019, 2004.
109. Malczewski, J., *GIS and multicriteria decision analysis*. John Wiley & Sons: New York, 1999.
110. Zionts, S., MCDM-If not a Roman Numeral, then what? *Interfaces.* 9(4):94–101, 1979.
111. Baltussen, R., and Niessen, L., Priority setting of health interventions: The need for multi-criteria decision analysis. *Cost Eff Resour Allocation.* 4(1):1, 2006.
112. Thokala, P., 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):1–13, 2016.
113. Oliveira, M., Fontes, D. B., and Pereira, T., Multicriteria decision making: A case study in the automobile industry. *Ann. Manag. Sci.* 3(1):109, 2014.
114. Whaiduzzaman, M., et al., Cloud service selection using multicriteria decision analysis. *Sci. World J.*, 2014.
115. Hwang, C., and Yoon, K., *Multiple Attribute Decision Making: Methods and Applications: A State-of-the-art Survey.Lecture Notes in Economics and Mathematical Systems*. Springer, 1981.
116. Aruldoss, M., Lakshmi, T. M., and Venkatesan, V. P., A survey on multi criteria decision making methods and its applications. *Am. J. Inf. Syst.* 1(1):31–43, 2013.

117. Guo, S., and Zhao, H., Fuzzy best-worst multi-criteria decision-making method and its applications. *Knowl.-Based Syst.* 121(Supplement C):23–31, 2017.
118. Rezaei, J., Best-worst multi-criteria decision-making method. *Omega.* 53(Supplement C):49–57, 2015.
119. Tavana, M., and Hatami-Marbini, A., A group AHP-TOPSIS framework for human spaceflight mission planning at NASA. *Expert Syst. Appl.* 38(11):13588–13603, 2011.
120. Jumaah, F. M., Zaidan, A. A., Zaidan, B. B., Bahbibi, R., Qahtan, M. Y., and Sali, A., 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, 2017.
121. Azeez, D. et al., Comparison of adaptive neuro-fuzzy inference system and artificial neutral networks model to categorize patients in the emergency department. *SpringerPlus.* 2(1):416, 2013.
122. Ashour, O. M., and Okudan, G. E., Fuzzy AHP and utility theory based patient sorting in emergency departments. *Int. J. Collab. Enterp.* 1(3–4):332–358, 2010.
123. Salman, O., et al., 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.
124. Mills, A. F., A simple yet effective decision support policy for mass-casualty triage. *Eur. J. Oper. Res.* 253(3):734–745, 2016.
125. Adunlin, G., Diaby, V., and Xiao, H., Application of multicriteria decision analysis in health care: A systematic review and bibliometric analysis. *Health Expect.* 18(6):1894–1905, 2015.
126. Jumaah, F., 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.* 2018.
127. 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, 2017.
128. Zaidan, B., et al., A new approach based on multi-dimensional evaluation and benchmarking for data hiding techniques. *Int. J. Inf. Technol. Decis. Mak.* :1–42, 2017.
129. 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. *J. Circuits Syst. Comput.* 26(07):1750116, 2017.
130. Abdullateef, B. N., et al., An evaluation and selection problems of OSS-LMS packages. *SpringerPlus.* 5(1):248, 2016.
131. Opricovic, S., and Tzeng, G. -H., Compromise solution by MCDM methods: A comparative analysis of VIKOR and TOPSIS. *Eur. J. Oper. Res.* 156(2):445–455, 2004.
132. Zaidan, A., et al., Evaluation and selection of open-source EMR software packages based on integrated AHP and TOPSIS. *J. Biomed. Inform.* 53:390–404, 2015.
133. Nilsson, H., Nordström, E. -M., and Öhman, K., Decision support for participatory forest planning using AHP and TOPSIS. *Forests.* 7(5):100, 2016.
134. Singh, A., Major MCDM Techniques and their application-A Review. 4:15–25, 2014.
135. 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.
136. Yas, Q. M., Zadain, A. A., Zaidan, B. B., Lakulu, M. B., and Rahmatullah, B., 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(3): 1759002, 2017.
137. 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(7):2344–2372, 2018.
138. 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.* 2018.
139. Zaidan, 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.* :1–16, 2018.
140. Opricovic, S., and Tzeng, G. -H., Extended VIKOR method in comparison with outranking methods. *Eur. J. Oper. Res.* 178(2): 514–529, 2007.
141. Wan Ahmad, W. N. K., et al., 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.
142. 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.
143. Rezaei, J., Best-worst multi-criteria decision-making method. *Omega.* 53:49–57, 2015.
144. Rezaei, J., Best-worst multi-criteria decision-making method: Some properties and a linear model. *Omega.* 64(Supplement C): 126–130, 2016.
145. Zaidan, A. A., et al., Multi-criteria analysis for OS-EMR software selection problem: A comparative study. *Decis. Support. Syst.* 78(Supplement C):15–27, 2015.
146. Zavadskas, E. K., et al., Multi-Attribute Decision-Making Model by Applying Grey Numbers. *Informatica.* 20(2):305–320, 2009.
147. Medineckienė, M., and Bjork, F., Owner preferences regarding renovation measures - The demonstration of using multi-criteria decision making. 17:284–295, 2011.
148. 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'. *Software: Practice and Experience.* 47(10):1365–1392, 2017.
149. Mahjouri, M., et al., Optimal selection of Iron and Steel wastewater treatment technology using integrated multi-criteria decision-making techniques and fuzzy logic. *Process Saf. Environ. Prot.* 107(Supplement C):54–68, 2017.
150. Karahalios, H., The application of the AHP-TOPSIS for evaluating ballast water treatment systems by ship operators. *Transp. Res. Part D: Transp. Environ.* 52(Part A):172–184, 2017.
151. Behzadian, M., et al., A state-of the-art survey of TOPSIS applications. *Expert Syst. Appl.* 39(17):13051–13069, 2012.
152. Shih, H. -S., Shyur, H. -J., and Lee, E. S., An extension of TOPSIS for group decision making. *Math. Comput. Model.* 45(7):801–813, 2007.
153. Kaur, S., Sehra, S. K., and Sehra, S. S., A framework for software quality model selection using TOPSIS. In: *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. 2016.
154. Sutadian, A. D., et al., Using the Analytic Hierarchy Process to identify parameter weights for developing a water quality index. *Ecol. Indic.* 75(Supplement C):220–233, 2017.
155. Sofuoglu, M. A., and Orak, S., A Novel Hybrid Multi Criteria Decision Making Model: Application to Turning Operations. *Int. J. Intell. Syst. Appl. Eng.* 5(3):124–131, 2017.
156. Raviv, G., Shapira, A., and Fishbain, B., AHP-based analysis of the risk potential of safety incidents: Case study of cranes in the construction industry. *Saf. Sci.* 91(Supplement C):298–309, 2017.
157. Zhao, H., Guo, S., and Zhao, H., Comprehensive benefit evaluation of eco-industrial parks by employing the best-worst method

based on circular economy and sustainability. *Environ. Dev. Sustain.* :1–25, 2017.
158. Chou, S. -Y., Chang, Y. -H., and Shen, C. -Y., A fuzzy simple additive weighting system under group decision-making for facility location selection with objective/subjective attributes. *Eur. J. Oper. Res.* 189(1):132–145, 2008.
159. Jablonsky, J., MS Excel based Software Support Tools for Decision Problems with Multiple Criteria. *Procedia Econ. Financ.* 12(Supplement C):251–258, 2014.
160. Rezaei, J., Best-worst multi-criteria decision-making method: Some properties and a linear model. *Omega.* 64:126–130, 2016.
161. Lo, H. -W., et al., An integrated model for solving problems in green supplier selection and order allocation. *J. Clean. Prod.* 190: 339–352, 2018.
162. Yang, Q., et al., Evaluation and Classification of Overseas Talents in China Based on the BWM for Intuitionistic Relations. *Symmetry.* 8(11):137, 2016.
163. 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.* 2018.
164. Gupta, H., Evaluating service quality of airline industry using hybrid best worst method and VIKOR. *J. Air Transp. Manag.* 68:35–47, 2018.
165. Serrai, W., et al., An efficient approach for Web service selection. In: *2016 IEEE Symposium on Computers and Communication (ISCC)*, 2016.
166. 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.
167. Serrai, W., et al., Towards an efficient and a more accurate web service selection using MCDM methods. *J. Comput. Sci.* 22:253–267, 2017.
168. 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.
169. Guo, S., and Zhao, H., Fuzzy best-worst multi-criteria decision-making method and its applications. *Knowl.-Based Syst.* 121:23–31, 2017.
170. Aboutorab, H., et al., ZBWM: The Z-number extension of Best Worst Method and its application for supplier development. *Expert Syst. Appl.* 107:115–125, 2018.
171. Rezaei, J., van Roekel, W. S., and Tavasszy, L., Measuring the relative importance of the logistics performance index indicators using Best Worst Method. *Transp. Policy.* 68:158–169, 2018.
172. Salimi, N., and Rezaei, J., Evaluating firms' R&D performance using best worst method. *Eval. Program Plann.* 66:147–155, 2018.
173. 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.
174. Ou Yang, Y. -P., Shieh, H. -M., and Tzeng, G. -H., A VIKOR technique based on DEMATEL and ANP for information security risk control assessment. *Inf. Sci.* 232:482–500, 2013.
175. Jahan, A., et al., A comprehensive VIKOR method for material selection. *Mater. Des.* 32(3):1215–1221, 2011.
176. Cavallini, C., et al., Integral aided method for material selection based on quality function deployment and comprehensive VIKOR algorithm. *Mater. Des.* 47:27–34, 2013.
177. Liou, J. J. H., et al., A modified VIKOR multiple-criteria decision method for improving domestic airlines service quality. *J. Air Transp. Manag.* 17(2):57–61, 2011.
178. Shojaei, P., Haeri, S. A. S., 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.
179. Golub, T. R., et al., Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science.* 286(5439):531–537, 1999.
180. Huang, P. H., and Moh, T. -t., A non-linear non-weight method for multi-criteria decision making. *Ann. Oper. Res.* 248(1):239–251, 2017.
181. 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.
182. Qader, M. A., et al., A methodology for football players selection problem based on multi-measurements criteria analysis. *Measurement.* 111:38–50, 2017.
183. Zaidan, A. A., et al., A survey on communication components for IoT-based technologies in smart homes. *Telecommun. Syst.* :1–25, 2018.
184. Alaa, M., Zaidan, A. A., Zaidan, B. B., Talal, M., and Kiah, M. L. M., A review of smart home applications based on Internet of Things. *J. Netw. Comput. Appl.* 97:48–65, 2017.
185. Zaidan, A. A., Karim, H. A., Ahmad, N. N., Alam, G. M., and Zaidan, B. B., A new hybrid module for skin detector using fuzzy inference system structure and explicit rules. *Int. J. Phys. Sci.* 5(13):2084–2097, 2010.
186. Zaidan, A. A., Karim, H. A., Ahmad, N. N., Zaidan, B. B., and Sali, A., An automated anti-pornography system using a skin detector based on artificial intelligence: A review. *Int. J. Pattern Recognit. Artif. Intell.* 27(04):1350012, 2013.
187. Zaidan, A. A., Abdul Karim, H., Ahmad, N. N., Zaidan, B. B., and Sali, A., A four-phases methodology to propose anti-pornography system based on neural and bayesian methods of artificial intelligence. *Int. J. Pattern Recognit. Artif. Intell.* 28(01):1459001, 2014.
188. Yas, Q. M., Zaidan, A. A., Zaidan, B. B., Hashim, M., and Lim, C. K., A systematic review on smartphone skin cancer apps: Coherent taxonomy, motivations, open challenges and recommendations, and new research direction. *J. Circuits, Syst. Comput.* 1830003, 2017.
189. Zaidan, B. B., Haiqi, A., Zaidan, A. A., Abdulnabi, M., Mat Kiah, M. L., and Muzamel, H., A security framework for nationwide health information exchange based on telehealth strategy. *J. Med. Syst.* 39(5):1–19, 2015.
190. Kiah, M. L. M., Haiqi, A., Zaidan, B. B., and Zaidan, A. A., Open source EMR software: Profiling, insights and hands-on analysis. *Comput. Methods Prog. Biomed.* 117(2):360–382, 2014.
191. Salman, O. H., Zaidan, A. A., Zaidan, B. B., Kalid, N., 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(5):1211–1245, 2017.
192. Kalid, N., Zaidan, A. A., Zaidan, B. B., Salman, O. H., Hashim, M., and Muzammil, H., 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):69, 2018.
193. Zaidan, A.A., Zaidan, B. B., Al-Haiqi, A, Kiah, M. L. M., Hussain, M. Evaluation and selection of opensource EMR software packages. *Elsevier.* 53, 2014.
194. Alanazi, H. O., Zaidan, A. A., Zaidan, B. B., Mat Kiah, M. L., and Al-Bakri, S. H., Meeting the security requirements of electronic medical records in the ERA of high-speed computing. *J. Med. Syst.* 39(1):1–14, 2015.
195. Alanazi, H. O., Alam, G. M., Zaidan, B. B., and Zaidan, A. A., Securing electronic medical records transmissions over unsecured

communications: An overview for better medical governance. *J. Med. Plant Res.* 4(19):2059–2074, 2010.

196. Hussain, M., Al-Haiqi, A., Zaidan, A., Zaidan, B., Kiah, M. L. M., Anuar, N. B., and Abdulnabi, M., The landscape of research on smartphone medical apps: Coherent taxonomy, motivations, open challenges and recommendations. *Comput. Methods Prog. Biomed.* 122(3):393–408, 2015.
197. Mat Kiah, M. L., Zaidan, B. B., Zaidan, A. A., Nabi, M., and Ibraheem, R., MIRASS: Medical informatics research activity support system using information mashup network. *J. Med. Syst.* 38(4):1–37, 2014a.
198. Mat Kiah, M. L., Al-Bakri, S. H., Zaidan, A. A., Zaidan, B. B., and Hussain, M., Design and develop a video conferencing framework for real-time telemedicine applications using secure group-based communication architecture. *J. Med. Syst.* 38(10):1–13, 2014c.
199. Mat Kiah, M. L., Nabi, M. S., Zaidan, B. B., and Zaidan, A. A., An enhanced security solution for electronic medical records based on AES hybrid technique with SOAP/XML and SHA-1. *J. Med. Syst.* 37(5):1–16, 2013.
200. Abdulnabi, M., Al-Haiqi, A., Kiah, M. L. M., Zaidan, A. A., Zaidan, B. B., and Hussain, M., A distributed framework for health information exchange using smartphone technologies. *J. Biomed. Inform.* 69:230–250, 2017.
201. Zaidan, B. B., Zaidan, A. A., and Mat Kiah, M. L., Impact of data privacy and confidentiality on developing telemedicine applications: A review participates opinion and expert concerns. *Int. J. Pharm.* 7(3):382–387, 2011.
202. 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):1–14, 2015.
203. Hussain, M., Ahmed, A.-H., Zaidan, A. A., and Zaidan, B. B., M Kiah, Salman Iqbal, S Iqbal, Mohamed Abdulnabi "A security framework for mHealth apps on Android platform". *Comput. Secur.* 45:191–217, 2018.
204. Hussain, M., Zaidan, A. A., Zidan, B. B., Iqbal, S., Ahmed, M. M., Albahri, O. S., and Albahri, A. S., Conceptual Framework for the Security of Mobile Health Applications on Android Platform. 35(3):1–32. 2018.
205. Albahri, A. S., Zaidan, A. A., Albahri, O. S., Zaidan, B. B., and Alsalem, M. A., Real-Time Fault-Tolerant mHealth System: Comprehensive Review of Healthcare Services, Opens Issues, Challenges and Methodological Aspects. *J. Med. Syst.* 42(8): 137, 2018.
206. Albahri, O. S., Zaidan, A. A., Zaidan, B. B., Hashim, M., Albahri, A. S., and Alsalem, M. A., 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.
207. 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.
208. 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):80, 2018.
209. Nabi, M. S. A., Kiah, M. M., Zaidan, B. B., Zaidan, A. A., and Alam, G. M., Suitability of using SOAP protocol to secure electronic medical record databases transmission. *Int. J. Pharmacol.* 6(6):959–964, 2010.
210. Rahmatullah, B., Zaidan, A. A., Mohamed, F., and Sali, A., 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, pp. 1084–1088, 2017. https://doi.org/10.1109/CoDIT.2017.8102743
211. Hussain, M., Al-Haiqi, A., Zaidan, A. A., Zaidan, B. B., Kiah, M. M., Anuar, N. B., and Abdulnabi, M., The rise of keyloggers on smartphones: A survey and insight into motion-based tap inference attacks. *Pervasive Mob. Comput.* 25:1–25, 2016.