Journal of Theoretical and Applied Information Technology
30th November 2016. Vol.93. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AUTOMATIC QUESTION CLASSIFICATION MODELS FOR COMPUTER PROGRAMMING EXAMINATION: A SYSTEMATIC LITERATURE REVIEW

[1] MUSTAFA KADHIM TAQI, [2] ROSMAH ALI
[1,2] Advanced Informatics School, Universiti Teknologi Malaysia (UTM), 54100, Kuala Lumpur, Malaysia
E-mail: [1] ktmustafa2@live.utm.my, [2] rosmaha.kl@utm.my

ABSTRACT

A test is the commonest method to evaluate the progress and potential of candidates in any field, especially in academic fields. In the academic field, exams help evaluate the understanding, applicative ability, and retainable knowledge of a student. Therefore, the questions need to be suitably set, so that all these areas can be judged. The results of these tests help determine if the student is fit for the next level of education. Setting the right question paper is a major challenge for the teachers concerned. The current study aims to analyze the ongoing question classification models with reference to the set of formulated research questions. In order to locate question classification models, relevant keywords were used in the search terms. A set of nine different studies were picked from the search processes. In the studies, 4 stands for journal articles, and 5 stands for conference papers. Question classification has been discussed in the computing domain, especially with respect to computer programming assessment. A more extensive examination of this classification reveals quite a few shortcomings of the prevailing classification methods. These include the absence of suitable taxonomy for computer programming questions, limitation in approaches to handle multi-labelling, and a need for methods compatible to tackle code-mixed question classification. Furthermore, the necessity to develop advanced hybrid feature selection methods in order to enhance the classification performance.

Keywords: *Question Classification, Feature Selection, Bloom Taxonomy, Computer Programming, Systematic Literature Review*

1. INTRODUCTION

For any IT student, introductory programming courses are necessary. It helps build the foundational concepts of programming, which are common to most of the core IT courses, refer to Sahami, Roach [1]. End of course formal examinations are one of the main techniques used for summative assessment of students in programming courses.

Construction of an examination instrument is an important task, as the exam is used both to measure the level of knowledge and skill that students have reached at the end of the course and to grade and rank the students. A poorly constructed exam may not give a fair assessment of students' abilities, perhaps affecting their grades and their progression through their program of study.

Developing an examination paper is often an individual task, with the exam's format depending on the examiner's own preferences as well as on examination questions inherited from colleagues in previous offerings of the course. There are a number of ways that skills and knowledge may be assessed, and exams typically have a number of questions in a variety of styles, giving students different ways to demonstrate their knowledge and skills.

In constructing an exam, educators must consider what they wish to assess in terms of the course content. They must consider the expected standards of their course and decide upon the level of difficulty of the questions.

Tew [2] claims that "the field of computing lacks valid and reliable assessment instruments for pedagogical or research purposes". This is a concern, because if the instruments we are using are neither valid nor reliable, how can we rely upon our interpretation of the results? [3].

Considering the central role of the formal examination in assessing our students, it is important to ensure that the questions are balanced for low and high difficulty levels, and to ensure an effective pattern of questions that will aid optimum learning in students. According to Swart [4], the educators setting the question need to follow some

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

classification guidelines such as a learning taxonomies. However, this seems not a straightforward task especially to newbies on computing teaching or those educators whose English is not their first language. Therefore, several automatic classification models have been developed. These models categorize any question in its corresponding difficulty level. For example, [5-12].

Even though there are a considerable number of publications regarding question classification techniques, little attention has been paid to analyse them in a systematic way. Most recently, Anbuselvan, Manoranjitham [13] presented a complete literature survey of current methods or approaches for question classification. However, they limited their review only to statistical approaches. They looked at literature from two viewpoints. First, the domain of application that consists question answering systems, information retrieval and educational environment. Second, the languages used in these classification methods. They identified several question classification methods that used to categorize different languages include English, Chinese, Dutch, Italian, and Spanish.

The scope of the current paper, however, concentrates on the educational domain by assessing the present status of Question Classification Models (QC) post 2010. In fact, this study aims to highlight the importance and challenges of the research of question classification for computer programming. The objective is to help researchers in this field to understand the whole picture of the current research in this topic, and facilitate them to choose suitable techniques in their research.

Furthermore, this paper expands the previous work in several ways. Firstly, a systematic and comprehensive survey is provided by including more question classification techniques, especially some very recent techniques not mentioned in Anbuselvan, Manoranjitham [13]. Secondly, the trends and open questions in this research topic is discussed and some guidelines for the future research are proposed.

The criteria used for evaluating the reviewed literature are based on the classification approaches used for question categorisation, Schemas used in the categorisation, the feature selection methods, the evaluate method, the types of classification, and the Language of the existing questions set.

Here are the different sections the study has been categorized into: Section 2 provide relevant background information. Section 3 present a description of the methodology used in this study. The presentation of results and its discussion are in Section 4 and 5 respectively. In Section 6, the issues and implications are discussed. The conclusion presented in Section 7.

2. BACKGROUND

2.1 Question Classification and its Approaches

Text classification is the task of discovering the category or categories that a text document belongs to, from a fixed set of predefined categories. In other words, to assign category labels to documents. Question classification is the method which is used for evaluating a question and labeling it on the basis of the estimated answer type [14]. The fixed set of estimated answer types are recognized as question ontology or taxonomy or category, the main objective of the question classification system is to find out a map on the basis of questions to answer types. Although the process may look very easy, and can of course be done manually, but in this paper we look at automated question classification systems.

The working procedure of the question classification is similar to the document classification but document classification helps in achieving more accuracy in classifying question compare to that of question classification. This occurs due to the document classification consist of more information or words which classify questions as compare to question classification [15]. The result will have a great effect on the perceptive power in classifying question.

2.1.1 Rule-based question classification

The rule-based question categorizes the problems in a simple way through the use of pre-configured heuristic rules which is mainly based on taxonomy. The experts use the rule-based techniques for organizing the question on the basis of crafted rules.

A rule dependent categorizer is accessible and estimated in Hovy, Hermjakob [16]. The classifier mainly uses the rules for detecting the question headword and also makes use of Word Net for mapping the target category. In this process, machine learning categorizer utilizes the attributes of a rule based classifier for getting into the concluding steps.

A new and fused pattern which is mainly based on rule based classifier and statistical process is projected in May and Steinberg [17]. The QC is planned through the use of Markov logic network and utilizes a blurred discriminative learning

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

process. The rule based process is not sustainable on other fields or in various languages due to its difficulty in framing a novel set of rules. The rule-based approaches execute better in a certain dataset and compare to the condensed concert on the fresh dataset. The rule based technique is precise in forecasting a positive group of questions; although, it is not measurable to a huge amount of queries and syntactical arrangements.

2.1.2 Machine learning question classification

Machine learning is a form of synthetic intellect which makes the processor with the capability to become skilled devoid of being openly planned. Machine learning mainly centered on the improvement of computer programs which can help them in teaching themselves in growing and changing. Machine Learning (ML) has become extremely popular in the current times as it is performing extensive assortment of vital functions, such as data mining, image recognition, natural language processing, and expert systems. It also offers possible solutions in all these domains and more. It is mainly classified into two types:

2.1.2.1 Supervised machine learning

Supervised Machine Learning is "trained" on a pre-configured set of "training examples" which has the capacity to achieve a precise ending at the time of giving new date. It is quite regular in categorization problems. Supervised learning is the widely used approach for guiding neural networks and decision trees. The technique is mainly based on the information provided by the programmed categorizations. The technique mainly resolves the network errors and then regulates the network for minimizing it, and in decision trees, it is mainly used to find out the characteristics that provides the most essential data which can be utilized to explain the cataloging puzzle.

2.1.2.2 Unsupervised machine learning

Unsupervised learning is much harder as compare to others: the objective is to make computer learn how to perform some tasks which is told to do. There are mainly two aspects in unsupervised learning.

The first approach is to instruct the means by not giving precise, but by means of reward system to specify accomplishment. In this context, these kinds of exercise will usually synchronize into the concluding error structure because the objective is not to create a categorization however to build conclusion that make best use of rewards.

A second form of unsupervised learning is known as clustering. In this sort of learning, the objective is not to enhance the effectiveness of the function, but generally to un cover the correspondences in the training information. The hypothesis is frequent because of the fact that the clusters revealed will synchronize sensibly fit with a spontaneous categorization. The clustering of individuals is based on statistical data which might result in a clustering of affluent in one cluster and the reduced in another cluster.

ML method based QC employs, syntactic, lexical and semantic aspects. An Extreme Learning Machine (ELM) utilizes semantic characteristics to enhance equally the training and testing in contrast to the standard of the SVM classifier [18]. The purpose of ELM is to categorize the semantic aspect of arithmetical QC.

2.1.2.3 Hybrid approach

The hybrid technique coalesces the characteristic set of two or further QC techniques. Hybrid process connect the conception of rule-based and learning based system, it anticipated the hybrid process that utilize the data of headwords and groups from rule-based classifier to create attributes set for guidance and combine this information with the data attained from the question unigrams [16].

Question classification is a main issue of Community Question Answering (CQA) services [19]. It is not easy to learn the concert of the advanced emerging processes in QC or short text classification. It unites Support Vector Machine (SVM), Naïve Bayes (NB) and Maximum Entropy (ME) techniques by means of different trait demonstration in some section of QC.

2.2 Bloom's Taxonomy

In general, Bloom's Taxonomy is a multi-level classification model that classifies thinking according to cognitive complexity. Any educational taxonomy should be able to classify the intended student behaviors (mental behaviors) [20]. The taxonomy can be viewed as a one-dimensional continuum model [20, 21] or as a matrix as in the revised Bloom's Taxonomy (RBT) [22]. One of the main purposes of creating taxonomy of educational objectives is to facilitate communication [20], i.e., to provide a common language for defining intended learning outcomes and student performance in assessments [23].

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Obviously, the Bloom taxonomy, developed by Bloom, Engelhart [20], plays a significant role in education as a generic tool for separating the cognitive features of learning into a hierarchy comprising six echelons (see Figure 1). However, Bloom, Engelhart [20] were rather unclear about whether "evaluation" should be placed above or on the same tier as "synthesis", and they were also uncertain whether an elevated achievement at a higher tier automatically reflects performance on the lower tiers.

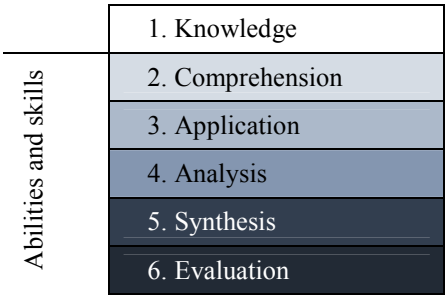

Figure 1: Bloom's Taxonomy [20]

Over forty years later, Anderson, Krathwohl [22] revised Bloom's Taxonomy by changing the nouns in Bloom's version to verbs to make the taxonomy consistent with how learning goals are characteristically expressed. Anderson, Krathwohl [22] believe that a taxonomy should be a two-dimensional approach. According to them, the cognitive process dimension should comprise the fundamental classifications of "remember", "understand", "apply", "analyze", "evaluate" and "create", whereas the information element should comprise the classifications of "factual", "conceptual", "procedural" and "meta-cognitive" (see Figure 2).

There is a clear-cut disparity between the revised and original versions of the taxonomy. A learner performing at a higher tier will be able to match this performance at lower tiers in the revised version. Moreover, this model appears to favor a chronological learning process but does not completely dismiss an iterative route to learning.

2.3 Feature Selection

A feature is some characteristic, detail or aspect of something. In a text document, the words or terms make up the most obvious features, but as we shall see in the next section, other ways of finding features exist as well. For the task of classifying questions, we need good quality features. Good quality features contain much information that the classifier can use to decide which category a question belongs to.

	Factual	Conceptual	Procedural	Metacognitive
1. Remember				
2. Understand				
3. Apply				
4. Analyze				
5. Evaluate				
6. Create				

Figure 2: Revised Bloom's Taxonomy [22]

Hence, a word that occurs in all questions in one category in the training set and in none of the other categories, would probably be a good quality feature for that category. Poor quality features contain less information about the class membership. For instance, stop words (like "the", "it", "and" etc.) will probably occur in all categories, and would not help the classifier much in the decision process. Poor quality features are also called irrelevant features [24], and the performance of the classifier is maintained (or even raised) if they are removed.

Therefore, Feature Selection is the step of selecting some features (e.g. words or terms) to be used when building an automatic classifier for question classification. Special kind of irrelevant features are the redundant features. They are useful for the classifier themselves, but can be removed since there are other features contributing the same information.

Note that by removing redundant features, the classifier performance remains the same, while the computation time falls. Some features can seem like good features in the training data, but then turn out not to work well in real life. Such features are called noise features, and when such features are selected by the feature selection technique, it is called over fitting, i.e. the classifier trained with the selected

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

features (including the noise features) will be very good at categorizing the training documents, but not so good for other documents. As pointed out in [25] there are two main reasons for selecting some features over others:

Accuracy- Firstly, studies have shown that machine learning algorithms can produce better results when not considering all the features. It would be reasonable to think that the more features considered, the more accurate the classifier would become. However, some features do not add more information (they are merely noise), and removing these can make the classifier perform more accurately.

Scalability-Secondly, as machine learning algorithms are resource demanding (computation power, memory need, network bandwidth, storage, etc.), running them on a subset of the features typically yields significant time savings. The ability to work with a small subset of the features also ensures scalability. Mladenić [26] reported good accuracy even with subsets of just 2% of the available features.

Feature Selection methods are often grouped into filters and wrappers. Filter methods measure feature relevance by applying statistical tests to the feature counts. Wrapper methods measure feature subset usefulness by using Where filter methods evaluate each feature independently, wrappers evaluate feature sets as a whole, which in theory would avoid redundant features and lead to better results. However, wrapper methods are computationally infeasible for large datasets, and are also more prone to overfitting, so filter methods are more commonly used. The following are description on some currently used filter approaches to feature selection.

However, we should introduce some notations to be used in the description. The F is for Feature. Categories C_k are labels. N is the total number of documents in the training set. N_{C_k} is the number of documents in category C_k. $N_{\overline{C_k}}$ is the number of documents not in category C_k. N_F is the number of documents containing feature F. $N_{\bar{F}}$ is the number of documents not containing feature F. N_{F,C_k} is the number of documents containing feature F in category C_k. $N_{\bar{F},C_k}$ is the number of documents not containing feature F in category C_k. $N_{F,\overline{C_k}}$ is the number of documents containing feature F not in category C_k. $N_{\bar{F},\overline{C_k}}$ is the number of documents not containing feature F not in category C_k.

2.3.1 Term Frequency (TF)

The simple Term Frequency for a (feature, category) pair is defined by Eyheramendy and Madigan [27] as the number of documents in category C_k containing feature F, as shown in Equation (1). Hence it looks only for positive evidence of category membership.

$$TF(F, C_k) = N_{F,C_k} \tag{1}$$

As done in Eyheramendy and Madigan [27], we aggregate these values to find the global Term Frequency value for each feature F by weighting the value of each (feature, category) pair by the category dominance and then summarize the weighted values:

$$TF(F) = \sum_{k=1}^{|C|} \frac{N_{C_k}}{N} N_{F,C_k} \tag{2}$$

2.3.2 Odds Ratio (OR)

Odds Ratio [26, 28, 29]evaluates the probability of a traits taking place in one group with the chances for it happening in different group. It provides a positive gain to characteristics that take place more frequently in one group than in the other category, and a negative score arises if in more in the other. A score of zero indicates the odds for a attribute to come about in one group is accurately the same as the odds for it to happen in the other, $ln\,(1) = 0$. The original Odds Ratio algorithm for binary:

$$\begin{aligned} \mathrm{OR}(\mathrm{F}, C_k) &= \ln \frac{P(F|C_k)\left(1 - P(F|C_k)\right)}{P(F|\overline{C_k})\left(1 - P(F|\overline{C_k})\right)} \\ &= \ln \frac{(\frac{N_{F,C_k}}{N_{C_k}})(1 - \frac{N_{F,\overline{C_k}}}{N_{\overline{C_k}}})}{(\frac{N_{F,\overline{C_k}}}{N_{\overline{C_k}}})(1 - \frac{N_{F,C_k}}{N_{C_k}})} \end{aligned} \tag{3}$$

where F is a feature, C_k is the category of consideration, $P(F|C_k)$ is the prospect for the attributes F to occur in category C_k, and $P(F|\overline{C_k})$ is the probability for the feature F to occur in category $\overline{C_k}$.

To estimate the probabilities, we use the figure of guidance documents in groupC_k containing the feature F divided by the total number of training documents in categoryC_k, and similarly for category$\overline{C_k}$.

$$P(F|C_k) = \frac{N_{F,C_k}}{N_{C_k}} \tag{4}$$

Journal of Theoretical and Applied Information Technology
30th November 2016. Vol.93. No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

An alert reader will notice that there might occur divide-by-zero and $ln(0)$ problems when using this estimation technique with equation (3). We follow [28] and treat singularities as special cases: When $P(F|\overline{C_k}) = 0$ because none of the training documents in category C_k contain the feature F, we substitute $P(F|\overline{C_k})$ with $\frac{1}{N^2}$. Also, when $P(F|\overline{C_k}) = 1$ because all the training documents in category C_k contains the feature F, we substitute $P(F|\overline{C_k})$ with $1 - \frac{1}{N^2}$, where N is the number of documents in the whole corpus/collection. Thus the equation for estimating the probabilities including the special cases becomes this:

$$P(F|C_k) = \begin{cases} \frac{N_{F,C_k}}{N_{C_k}} \\ \frac{1}{N^2} \quad if\ N_{F,C_k} = 0 \\ 1 - \frac{1}{N^2} \quad if\ N_{F,C_k} = N_{C_k} \end{cases} \tag{5}$$

Taking the square of the corpus size into consideration ensures that low probabilities are well estimated in small corpora.

2.3.3 Information Gain (IG)

The basic idea behind IG is to find out how well each single feature separates the given data set. Information entropy is used to measure the uncertainty of the feature (e.g. term) and the dataset (e.g. a corpus of documents). The Information Gain of a feature is computed by Equation (6).

$$\begin{aligned} IG(Feature) = & \sum_{k=1}^{C} \frac{N_{C_k}}{N} \ln \frac{N_{C_k}}{N} \\ & + \frac{N_F}{N} \sum_{k=1}^{C} \frac{N_{F,C_k}}{N_F} \ln \frac{N_{F,C_k}}{N_F} \\ & + \frac{N_{\bar{F}}}{N} \sum_{k=1}^{C} \frac{N_{\bar{F},C_k}}{N_{\bar{F}}} \ln \frac{N_{\bar{F},C_k}}{N_{\bar{F}}} \end{aligned} \tag{6}$$

Equation (6) takes the overall entropy for the training set (the first line of the equation) minus the entropy for the feature (the last two lines of the equation). In Equation (6), we calculate the expected reduction in entropy if we categorize the corpus according to that feature. After computing IG values for all features, we can use the features with the highest IG score as features in any text categorization classifier.

Notice that the overall entropy for the training set naturally is the same for all features. Hence Equation (6) could easily have been split up into the calculation of the overall entropy $H(Set)$ and the feature entropy values $H(Feature)$, and then for each feature the calculation of the Information Gain value $IG(Feature) = H(Set) - H(Feature)$.

2.3.4 Mutual Information (MI)

Mutual Information is a verified equivalent to attain data for binary problems. For diverse group issues however, the two are not equivalent. Therefore, we provide Mutual Information with its possessing equation as a divide characteristic assortment algorithm here. We compute the Mutual Information of a term and category pair as shown in Equation (7):

$$\begin{aligned} & MI(F, C_k) \\ & = \sum_{v_f \in (1,0)} \sum_{v_{C_k} \in (1,0)} P(F = v_f, C_k \\ & = v_{C_k}) \ln \frac{P(F = v_f, C_k = v_{C_k})}{P(F = v_f) P(C_k = v_{C_k})} \end{aligned} \tag{7}$$

where F is the separate arbitrary inconsistent "feature" that takes the value $v_F = (1,0)$ (feature F occurs in document or not), C_k is the distinct arbitrary variable "category" that takes the values $v_{C_k} = (1,0)$ (document belongs to category C_k or not).

The prospects can be predicted by means of the different document tally from the training set. Using the notation mentioned in the beginning of section 2.3, we rewrite Equation (8) into Equation (9):

$$\begin{aligned} MI(F, C_k) = & \frac{N_{F,C_k}}{N} \ln \frac{N N_{F,C_k}}{N_F N_{C_k}} \\ & + \frac{N_{F,\overline{C_k}}}{N} \ln \frac{N N_{F,\overline{C_k}}}{N_F N_{\overline{C_k}}} \\ & + \frac{N_{\bar{F},C_k}}{N} \ln \frac{N N_{\bar{F},C_k}}{N_{\bar{F}} N_{C_k}} \\ & + \frac{N_{\bar{F},\overline{C_k}}}{N} \ln \frac{N N_{\bar{F},\overline{C_k}}}{N_{\bar{F}} N_{\overline{C_k}}} \end{aligned} \tag{8}$$

Then the values can be weighted and summarized to create a global ranked list of features:

$$MI(F) = \sum_{k=1}^{|C|} \frac{N_{C_k}}{N} MI(F, C_k) \tag{9}$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2.3.5 Chi Square (CHI)

Feature selection by x^2 testing [30, 31] is based on Pearson's x^2 (chi square) test. And the x^2 test is frequently applied for the assessment the self-rule of two inconsistent. The null-hypothesis is this case; the two variables are totally self-regulating on one another. The increase in the value of the x^2 analysis denotes the relationship between the variables have more closely. In feature selection, the x^2 test determines the self-rule of a trait and a category. The null-hypothesis in this context, the characteristic and group are totally autonomous, i.e. that the trait is ineffective for classifying documents.

$$x^2(F, C_k) = \frac{N \times \left(\left(N_{F,C_k} \times N_{\bar{F},\overline{C_k}}\right) - \left(N_{F,\overline{C_k}} \times N_{\bar{F},C_k}\right)\right)^2}{N_F \times N_{\bar{F}} \times N_{C_k} \times N_{\overline{C_k}}} \quad (10)$$

2.4 Classification Evaluate Methods

Automatic question classification systems are estimated based on how sound they execute the evaluation to the exact group information of questions. Utilizing questions with known category information, we can evaluate a classifiers concert in various metrics:

The proportion of properly categorized questions is a natural starting point for measuring the performance of a classifier. However, as some important information may be hidden behind this metric, a few other measures are commonly used as well. Before presenting these, we show some important question counts used in the computations. Table 1 represents the contingency table for evaluation measures.

	Question actually belongs to category	Question actually does not belong to category
Classifier says question belongs to category	True positives (TP)	False Positives (FP)
Classifier says question does not belong to category	False negatives (FN)	True negatives (TN)

Table 1: Notation: Contingency table for evaluation measures

Here, a true positive (TP) is a correctly assigned positive class label, while a false positive (FP) is an incorrectly assigned positive class label. Hence, the sum of true and false positives means the actual number of questions that the classifier placed in that class, while the sum of true positives and false negatives means the number of questions that actually belong to a category, no matter what the classifier says.

Hence, the sum of true and false positives means the actual number of questions that the classifier placed in that class, while the sum of true positives and false negatives means the number of questions that actually belong to a category, no matter what the classifier says.

The precision of a classifier is defined as the the number of true positives divided by the number of true and false positives. The recall of a classifier is defined as the the number of true positives divided by the number of true positives and false negatives [32].

$$\text{Precision} = \frac{TP}{TP + FP} \quad (11)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (12)$$

The precision level is sometimes referred to as a level of exactness, while the recall level measures completeness. Some applications are more concerned with one than the other. For instance, when filtering junk e-mail from important e-mail, we would typically rather accept occasional spam messages in the inbox than important messages in the junk mail folder. Hence, when classifying spam, we would like a high degree of exactness or precision, while when classifying relevant messages, we would like a high degree of completeness or recall.

The precision and recall levels are often combined into a single metric called the F1-measure (or sometimes just F-measure). The number one says that precision and recall are equally weighted.

$$F_1 = \frac{2(\text{precision} \cdot \text{recall})}{\text{precision} + \text{recall}} \quad (13)$$

The F_1-measure is in fact a special case of the F_β-measure :

$$F_\beta = \frac{(1 + \beta^2)(\text{precision} \cdot \text{recall})}{\beta^2 \cdot \text{precision} + \text{recall}} \quad (14)$$

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Where β specifies how many times more recall should be weighted over precision [33]. For instance, the F2-measure weights recall twice as much as precision, while the F0:5 -measure weights precision twice as much as recall. In our experiments, we report the percentage of correctly categorized questions, precision, recall, and F1-measure of the classifiers performances.

Moreover, when categorizing questions into multiple classes, the precision, recall and F-measures can be reported in both a macro-average and micro average metric. Macro-average means the arithmetic mean over all classes, while micro-average means the average weighted by the class distribution.

The macro-average weights each problem equally, while the micro-average weights each question classification equally. Hence, for highly skewed class sizes, these figures may show quite different values. In our experiments however, we have used a corpus with rather similar question counts for each class, so micro- and macro-average values are rather similar. Thus we present only the macro-averaged precision, recall and F1 -measure.

2.1. Types of Text Classification

Text classification (TC) systems can either have several categories to choose from, or just two (e.g. interesting or uninteresting). Also, they can label each question with either exactly one, or several (0-k) category labels. We can break down the various types as follows:

Multiple Label TC Some systems and applications can assign from none to multiple category labels (zero, one, several, or even all) to each question. Articles in an online newspaper could for instance belong to both the `international' category and the "sports" category. This type of text classification systems is also called overlapping categories, as mentioned in Sebastiani [34].

Single Label TC Single label systems assign exactly one category label to each question. There can be many categories to choose from, however.

Binary TC Binary text classification is a unique single label text classification, but here only two classifications are accessible. Moreover, each question has to be labeled with one of these categories. Posts in a news feed could for instance be labeled interesting or uninteresting for a user. A junk mail filter is another example where a binary text classification system could be applied.

Note that binary classification also is important because it is often used as a subroutine in many multi-class (i.e. multiple label and single label) tasks, see [25].

3. RESEARCH METHOD

The study is in the form of a systematic literature review with reference to the original guidelines in Kitchenham and Charters [35]. The aim is to evaluate the obtainable studies of the QC model in the preliminary programming examinations. The exact steps for this method are stated as follows.

3.1. Research questions

The research questions addressed by this study are:

- ***RQ1:*** What are the approaches used in the available classification systems for categorizing of Computer Programming Questions (PQC)?
- ***RQ2:*** What classification schemas are used in the available PQC classifiers?
- ***RQ3:*** What feature selection methods are used in the available PQC classifiers?
- ***RQ4:*** What class-labeling methods are used in the available PQC classifiers?
- ***RQ5:*** What criteria are used to evaluate the available PQC classifiers?
- ***RQ6:*** What benchmarks languages are used in the available PQC experiments?

3.2. Identification of Relevant Literature

The search strings are constructed following the strategy in Salleh, Mendes [36]. They are:

- Deriving major applied terms in the evaluation questions with reference to population, interference, result, and context.
- Recording known keywords in the article.
- Looking for synonyms or substitute words for the listed keywords.
- Applying Boolean OR for the incorporation of alternate spellings and words.
- Applying Boolean AND for linking major terms based on population, interference, and result.

It is stated in Petticrew and Roberts [37] that the most important concerns in carrying out an SLR search are:

- **Sensitivity of search**- Refers to the number of relevant studies recovered.
- **Specificity of search**- Refers to the restriction of irrelevant studies.

The search string for this went something like: ("programming" OR "computer programming" OR "introductory programming") AND ("question classification" OR "question categorization").

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

The preliminary search involved the application of six online databases known to index QC primary studies. They are ACM Digital library, IEEEXplore, Science-Direct, SpringerLink, Wiley Online Library, ebscohost. Khan, Kunz [38] states the importance of referring to multiple databases to gather enough citations to avoid biased review. Full articles need to be researched in order to avoid this bias.

Kitchenham and Charters [35] remark on the importance of SE researchers for the identification of relevant databases online to aid the search process. When the primary search phase is done, the identification of applicable literature moves on into the secondary search phase, where the references identified in the primary phase are reviewed. Suitable papers were listed under studies that are fit for the synthesis.

3.3. Inclusion and Exclusion Criteria

The inclusion criteria aim at the inclusion of the empirical studies of the QC, which target the PQC. The search covers the published studies dating within 2010-2016. The exclusion criteria include QC papers that do not target the QC. Other criteria applied includes:

- Papers that claim authorship without evidence.
- Papers that scrutinize the Question Answering.
- Papers written in other languages.

3.4. Data Extraction and Study Quality Assessment

A form devised to collect substantiation applicable to the research questions, would help the data extraction procedure, and also help judge the quality of the primary readings. Four questions were selected from the Salleh, Mendes [36] for the quality check (See Table 2).

Table 2: Inclusion and exclusion criteria.

No	Item	Answer
1	Was the article refereed?	Yes/No
2	Were the aim(s) of the study clearly stated?	Yes/No/Partially
3	Were the data collection carried out very well?	Yes/No/Partially
4	Were the findings credible? For example, the study was methodologically explained so that we can trust the findings	Yes/No/Partially

Four general questions were selected to evaluate the quality in ratio scale format where:

- Yes= 1 Point.
- No= 0 Point.
- Partially = 0.5 Point.

The possible outcomes range from 0.0-4.0, 0.0 being the poorest, and 4.0 being the best quality.

One author was responsible for the completion of the form for all the primary studies. The first and the second authors validated the data extraction procedure, which was then compared in a meeting to discuss the review. For every point where the extracted data differed by less than 10-15 percent, an agreement was reached through dialogue. Since the aim was to reach an absolute conclusion on the sample, no inter-rater was measured [39].

4. RESULTS

4.1. Search Results

In this section, the results of the study are laid down. Table 3 shows the results of the search procedure. Three filters were used to select the most relevant studies, which are title, abstract, and full text scanning. It can be seen that the total studies that gained from the search procedure were 278 papers from all databases under consideration. After filtering by title the remained papers were 32. The abstract scanning procedure resulting in 16 papers. While the final full text scanning retains 9 relevant studies presented in Table 4.

Table 3: Search procedure results.

Database	Total	Selected by Title	Selected by Abstract	Selected by full text
IEEEXplore	14	6	5	5
ACM Digital library	106	0	0	0
Science-Direct	86	24	9	2
SpringerLink	28	0	0	0
Wiley Online Library	42	0	0	0
ebscohost	2	2	2	2
Total	**278**	**32**	**16**	**9**

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 4: Selected Articles.

ID	Reference			Title
	Author	Year	Publisher	
S1	Jayakodi, Bandara [40]	2015	IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)	An automatic classifier for exam questions in Engineering: A process for Bloom's taxonomy
S2	Abduljabbar and Omar [6]	2015	Journal of Theoretical and Applied Information Technology	Exam questions classification based on Bloom's taxonomy cognitive level using classifiers combination
S3	Yahya and Osman [41]	2014	2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA)	Classification of high dimensional Educational Data using Particle Swarm Classification
S4	Yahya, Osman [42]	2013	Procedia - Social and Behavioral Sciences	Analyzing the Cognitive Level of Classroom Questions Using Machine Learning Techniques
S5	Yahya, Osman [43]	2013	2013 13th International Conference on Intelligent Systems Design and Applications	Educational data mining: A case study of teacher's classroom questions
S6	Haris and Omar [7]	2013	International Journal of Information Processing & Management	Determining Cognitive Category of Programming Question with Rule-based Approach
S7	Omar, Haris [10]	2012	Procedia-Social and Behavioral Sciences	Automated analysis of exam questions according to Bloom's taxonomy
S8	Haris and Omar [44]	2012	2012 7th International Conference on the Computing and Convergence Technology (ICCCT)	A rule-based approach in Bloom's Taxonomy question classification through natural language processing
S9	IYusof and Hui [45]	2010	10th International Conference on the Intelligent Systems Design and Applications (ISDA)	Determination of Bloom's cognitive level of question items using artificial neural network

4.2. Quality Evaluation

The quality studies were assessed using the criteria in Section 2.4. Table 3 shows the scores for the studies. Table 5 indicates in the last column, the number of questions in which researchers agreed. The results indicate that all the studies scored by 4.0.

Table 5: Quality evaluation.

Study	Q1	Q2	Q3	Q4	Total score	Initial rater agreement
S1	Y	Y	Y	Y	4	2
S2	Y	Y	Y	Y	4	2
S3	Y	Y	Y	Y	4	2
S4	Y	Y	Y	Y	4	2
S5	Y	Y	Y	Y	4	2
S6	Y	Y	Y	Y	4	2
S7	Y	Y	Y	Y	4	2
S8	Y	Y	Y	Y	4	2
S9	Y	Y	Y	Y	4	2

5. DISCUSSIONS

This section consists of a wholesome discussion of the obtained results. With respect to the classifiers used, the Rule-based classifier was the dominant (N=4). While, the chain of classifiers became second with 2 studies. In these studies, Support Vector Machine (SVM), Naïve Bayes (NB), and k-Nearest Neighbor (k-NN) have been used for building 3 classifiers. Particle Swarm Optimization (PSO), Bayes (NB), WordNet similarity, and Rochio Algorithm (RA) utilized only once. The same appearance is for all of The whole feature, the question frequency (DF), and the category frequency–question frequency (CF-DF).

The only schemas used in the process were Bloom Taxonomy (BT) and its revised version (RBT). However, the knowledge dimension for the RBT were omitted. While regarding the feature selection methods in the available PQC Classifiers, over 50

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

percent of the models did not use any feature selection method. The rest used the following:

- Mutual Information (MI)
- Chi-square (x^2)
- Term Frequency (TF)
- Information Gain (IG)
- Odd ratio (OR)

In the available PQC Classifiers, Single-labelling technique was used on all the classification models. While in order to evaluate the available PQC classifiers, several measure metrics were applied for the performance evaluation of the current classifiers. These include:

- Precision or Accuracy
- Convergence Time
- Convergence Error
- Recall and F^1

With respect to the language of the questions in the datasets used in the existing studies, a single language (English) was used. Table 6 summaries the findings of the study.

Table 6: Summary of Findings.

Study	QC Approach	QC Schema	Feature Selection	Types of Classification	Evaluate method	Benchmark Language
S1	Rule-based, WordNet similarity	RBT	None	Single-label	Accuracy	English
S2	Chain of Classifiers (SVM, NB, k-NN)	BT	MI, Odd ratio, X^2	Single-label	Precision, recall and F^1 measure metrics	English
S3	PSO	BT	TF	Single-label	Precision, recall and F^1 measure	English
S4	Chain of Classifiers (SVM, NB, k-NN)	BT	None	Single-label	Precision, recall, Accuracy, and F measure	English
S5	Chain of Classifiers (SVM, NB, k-NN, RA)	BT	TF, MI, IG, X^2	Single-label	Precision, recall and F^1 measure	English
S6	Rule-Based	BT	None	Single-label	Precision, recall and F^1 measure	English
S7	Rule-Based	BT	None	Single-label	Precision, recall and F^1 measure	English
S8	Rule-Based	BT	None	Single-label	Precision, recall and F^1 measure	English
S9	The whole feature, DF, and CF-DF	BT	None	Single-label	Precision, convergence time, and convergence error	English

SVM: Support Vector Machine, NB: Naïve Bayes, PSO: Particle Swarm Optimization, kNN: k-Nearest Neighbor, RA: Rochio Algorithm, DF: Document frequency, CF-DF: Category frequency–document frequency, TF: Term Frequency, MI: Mutual Information, IG: Information Gain, X^2: Chi Square.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

6. ISSUES AND IMPLICATIONS

In this paper, we have conducted an SLR for computer programming-related question classification models. However, literature reviewed suggest some issues for researchers to be considered in the future.

The reviewed literature indicate that the current classifiers used are mostly either Rule-based classifiers. Nonetheless, there is only one study discussed the use of a combination classifier model, i.e. the ensemble classifiers chain (ECC).

However, the rule-based approach uses rules determined manually by knowledgeable engineers, with the help of domain experts. Developing such rules is tedious and time consuming [13]. Rule-based systems lack portability and robustness abilities. Additionally, the high cost of rules maintenance goes up, even when data is only marginally altered. In addition, the use of rules is not effective when large data are employed since a large set of rules must be developed.

On the other hand, a common advantage of ensembles is their well-known effect of generally increasing overall predictive performance. Indeed, ECC for programming question classification (PQC) is in its infancy, and it is the author's hope that this paper will inspire researchers to bring ECC to its full potential.

With regard to the classification schemas, the meta-analysis results indicate the use of Bloom's Taxonomy (BT), and the revised version of the same, the RBT, which has two dimensions, the Cognitive dimension, and Knowledge dimension as presented in section 2.2. However, the RBT later dimension was omitted in the existing classification models for computer programming questions. This is despite of the existing belief of that learning objectives comprise from verb-noun frame, whereas a noun or a noun phrase refers to certain objective content and the verb or verb phrases describe the action meant for the content, or the cognitive process [46]. In addition, using BT or RBT has been recognized a complicated.

Fuller, Johnson [23] has acknowledged the inclusion of computer programming-incompatible terminologies and segments in both, BT and RBT. This misdirects the educators in the field of computer programming learning. Jayakodi, Bandara [5], Jayakodi, Bandara [40] recommends the examination of the efficiency of taxonomies other than BT for the classification of questions. Therefore, a more efficient taxonomy needs to be formulated. The new taxonomy should be based on pertinent language and rational programming related tasks needs to be developed.

While in order to offer a wholesome description of the constructive objectives of programming, the taxonomy must include the dimensions of types of knowledge, and cognitive processes. Related to this, existing single-labelling type of question classification cannot be used with a two dimensional taxonomy if each dimension treated as a standalone class label. Hence, multi-labelling becomes indispensable. This leads extending the use of existing evaluation method and utilize of multi-label evaluation metrics. More detail information on these metrics can be found in Zhang and Zhou [47].

Feature selection is often considered as a necessary preprocess step to analyze these data, as this method can reduce the dimensionality of the datasets and often conducts to better analyses, as discussed in section 2.3. While several feature selection techniques have been tried in the existing question classification models for computer programming, hybrid method that combine the existing filter and wrapper feature selector that may involves meta-heuristic algorithms have been omitted.

Recent studies have demonstrated that such algorithms efficiently converge to high-quality solutions for complex problems [48-50]. Therefore, researchers are highly encouraged developing meta-heuristics for feature selection that can help question classifiers getting good results. In fact, there are some outstanding yet unexplored algorithms such as the kidney-inspired search algorithm [51] need to have more treatment.

The study also stresses on the need to incorporate other languages like Arabic-English bilingual for different parts of the world. Pure English is not enough to reach out to all the active parts of the world. Hence, the development of a classification model which will be compatible with code-mixed question sets, is absolutely necessary.

7. CONCLUSION

For both IT educators and students, end of course formal examinations for introductory programming course are necessary. This summative assessment is used both to measure the level of knowledge and skill that students have reached at the end of the course and to grade and rank the students. Therefore, Construction of an examination instrument in such a way to ensure that the questions are balanced for low and high difficulty levels, and to ensure an effective pattern of questions that will aid optimum

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

learning in students is an important task. This is to give a fair assessment of students' abilities.

To do so, several automatic question classification models have been developed. In this paper, a systematic literature review on the existing question classification for computer programming has been done. Several classification approaches have been recognized. However, the performance of these approaches are still limited and a lot of improvement and employment to a more advanced approaches such as ECC are suggested.

Moreover, it has been recognized that only Bloom's Taxonomy (BT), and the revised version of the same, the RBT. However, numerous studies advocated that the use of this taxonomy in categorizing computer programming questions is not free of difficulties. Therefore, developing a new computer programming-related taxonomy with sufficient dimensions to categorize programming learning objectives and questions is highly needed. This proposition may lead to develop a multi-label classifier instead of the existing single class-labelling classification models.

The reviewed literature also asserted that existing single language classification models is not enough to reach out to all the active parts of the world. Therefore, the study stresses on the need to incorporate other languages like Arabic-English mixed-coded for different parts of the world. Accordingly, it is absolutely necessary to encourage the researchers to develop a classification models that will be compatible with code-mixed programming question sets.

Finally, the study recommends and highly encourages developing meta-heuristics for feature selection. This can help question classifiers getting good results.

REFRENCES:

[1] Sahami, M., et al., *ACM/IEEE-CS computer science curricula 2013: implementing the final report*, in *Proceedings of the 45th ACM technical symposium on Computer science education*. 2014, ACM: Atlanta, Georgia, USA. p. 175-176.

[2] Tew, A.E., *Assessing fundamental introductory computing concept knowledge in a language independent manner*. 2010.

[3] Sheard, J., et al. *Exploring programming assessment instruments: a classification scheme for examination questions*. in *Proceedings of the seventh international workshop on Computing education research*. 2011. ACM.

[4] Swart, A.J., *Evaluation of final examination papers in engineering: A case study using Bloom's Taxonomy.* Education, IEEE Transactions on, 2010. 53(2): p. 257-264.

[5] Jayakodi, K., et al., *WordNet and Cosine Similarity based Classifier of Exam Questions using Bloom's Taxonomy.* International Journal of Emerging Technologies in Learning, 2016. 11(4).

[6] Abduljabbar, D.A. and N. Omar, *Exam questions classification based on Bloom's taxonomy cognitive level using classifiers combination.* Journal of Theoretical and Applied Information Technology, 2015. 78(3): p. 447-455.

[7] Haris, S.S. and N. Omar, *Determining Cognitive Category of Programming Question with Rule-based Approach.* International Journal of Information Processing & Management, 2013. 4(3).

[8] Verdú, E., et al., *A genetic fuzzy expert system for automatic question classification in a competitive learning environment.* Expert Systems with Applications, 2012. 39(8): p. 7471-7478.

[9] Pillai, P.G. and J. Narayanan, *Question Categorization Using SVM Based on Different Term Weighting Methods.* International Journal on Computer Science and Engineering, 2012. 4(5): p. 938.

[10] Omar, N., et al., *Automated analysis of exam questions according to Bloom's taxonomy.* Procedia-Social and Behavioral Sciences, 2012. 59: p. 297-303.

[11] Ahmad, N.D., et al. *Automating preparation of exam questions: Exam Question Classification System (EQCS)*. in *2011 International Conference on Research and Innovation in Information Systems*. 2011.

[12] Zhang, D. and W.S. Lee, *Question classification using support vector machines*, in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*. 2003, ACM: Toronto, Canada. p. 26-32.

[13] Anbuselvan, S., M. Manoranjitham, and E.H. Lim, *Question classification using statistical approach: a complete review.* Journal of Theoretical and Applied Information Technology, 2015. 71(3).

[14] Metzler, D. and W.B. Croft, *Analysis of statistical question classification for fact-based questions.* Information Retrieval, 2005. 8(3): p. 481-504.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[15] Li, X. and D. Roth, *Learning question classifiers*, in *Proceedings of the 19th international conference on Computational linguistics - Volume 1*. 2002, Association for Computational Linguistics: Taipei, Taiwan. p. 1-7.

[16] Hovy, E., U. Hermjakob, and D. Ravichandran. *A question/answer typology with surface text patterns*. in *Proceedings of the second international conference on Human Language Technology Research*. 2002. Morgan Kaufmann Publishers Inc.

[17] May, R. and A. Steinberg, *Building a Question Classifier for a TREC-Style Question Answering System.* AL: The Stanford Natural Language Processing Group, Final Projects, 2004.

[18] Li, X. and D. Roth, *Learning question classifiers: the role of semantic information.* Natural Language Engineering, 2006. 12(03): p. 229-249.

[19] Wagstaff, K., *Machine learning that matters.* arXiv preprint arXiv:1206.4656, 2012.

[20] Bloom, B.S., et al., *Taxonomy of educational objectives: Handbook I: Cognitive domain.* New York: David McKay, 1956. 19: p. 56.

[21] Bower, M. *A taxonomy of task types in computing*. in *ACM SIGCSE Bulletin*. 2008. ACM.

[22] Anderson, L., D.R. Krathwohl, and B.S. Bloom, *A taxonomy for learning, teaching and assessing: a revision of Bloom's taxonomy of educational objectives*. 2001: Longman.

[23] Fuller, U., et al., *Developing a computer science-specific learning taxonomy.* ACM SIGCSE Bulletin, 2007. 39(4): p. 152-170.

[24] Liu, H. and H. Motoda, *Computational Methods of Feature Selection (Chapman & Hall/Crc Data Mining and Knowledge Discovery Series).* 2007.

[25] Forman, G., *Feature selection for text classification.* Computational methods of feature selection, 2008. 1944355797.

[26] Mladenić, D., *Feature subset selection in text-learning*, in *Machine Learning: ECML-98: 10th European Conference on Machine Learning Chemnitz, Germany, April 21–23, 1998 Proceedings*, C. Nédellec and C. Rouveirol, Editors. 1998, Springer Berlin Heidelberg: Berlin, Heidelberg. p. 95-100.

[27] Eyheramendy, S. and D. Madigan, *A bayesian feature selection score based on naive bayes models.* Liu and Motoda [LM07a], 2007: p. 277-294.

[28] Shaw, W.M., *Term-relevance computations and perfect retrieval performance.* Information Processing & Management, 1995. 31(4): p. 491-498.

[29] Chen, J., et al., *Feature selection for text classification with Naïve Bayes.* Expert Systems with Applications, 2009. 36(3, Part 1): p. 5432-5435.

[30] Yang, Y. and J.O. Pedersen. *A comparative study on feature selection in text categorization*. in *ICML*. 1997.

[31] Schütze, H. *Introduction to Information Retrieval*. in *Proceedings of the international communication of association for computing machinery conference*. 2008.

[32] Baeza-Yates, R. and B. Ribeiro-Neto, *Modern information retrieval*. Vol. 463. 1999: ACM press New York.

[33] van Rijsbergen, C., *Information Retrieval. 1979*. 1979, Butterworth.

[34] Sebastiani, F., *Machine learning in automated text categorization.* ACM Comput. Surv., 2002. 34(1): p. 1-47.

[35] Kitchenham, B.A. and S. Charters, *Guidelines for performing systematic literature reviews in software engineering.* 2007.

[36] Salleh, N., E. Mendes, and J. Grundy, *Empirical studies of pair programming for CS/SE teaching in higher education: A systematic literature review.* Software Engineering, IEEE Transactions on, 2011. 37(4): p. 509-525.

[37] Petticrew, M. and H. Roberts, *Systematic reviews in the social sciences: A practical guide*. 2008: John Wiley & Sons.

[38] Khan, K., et al., *Systematic reviews to support evidence-based medicine*. 2011: Crc Press.

[39] Madeyski, L., *On the effects of pair programming on thoroughness and fault-finding effectiveness of unit tests*, in *Product-Focused Software Process Improvement*. 2007, Springer. p. 207-221.

[40] Jayakodi, K., M. Bandara, and I. Perera. *An automatic classifier for exam questions in Engineering: A process for Bloom's taxonomy*. in *Teaching, Assessment, and Learning for Engineering (TALE), 2015 IEEE International Conference on*. 2015. IEEE.

[41] Yahya, A.A. and A. Osman. *Classification of high dimensional Educational Data using Particle Swarm Classification*. in *2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA)*. 2014.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[42] Yahya, A.A., et al., *Analyzing the Cognitive Level of Classroom Questions Using Machine Learning Techniques.* Procedia - Social and Behavioral Sciences, 2013. 97: p. 587-595.
[43] Yahya, A.A., A. Osman, and A.A. Alattab. *Educational data mining: A case study of teacher's classroom questions*. in *2013 13th International Conference on Intellient Systems Design and Applications*. 2013.
[44] Haris, S.S. and N. Omar. *A rule-based approach in Bloom's Taxonomy question classification through natural language processing*. in *Computing and Convergence Technology (ICCCT), 2012 7th International Conference on*. 2012.
[45] IYusof, N. and C.J. Hui. *Determination of Bloom's cognitive level of question items using artificial neural network*. in *Intelligent Systems Design and Applications (ISDA), 2010 10th International Conference on*. 2010. IEEE.
[46] Amer, A., *Reflections on Bloom's revised taxonomy.* Electronic Journal of Research in Educational Psychology, 2006. 4(1): p. 213-230.
[47] Zhang, M.L. and Z.H. Zhou, *A Review on Multi-Label Learning Algorithms.* IEEE Transactions on Knowledge and Data Engineering, 2014. 26(8): p. 1819-1837.
[48] Babatunde, O., et al., *A genetic algorithm-based feature selection.* British Journal of Mathematics & Computer Science, 2014. 4(21): p. 889-905.
[49] Xiang, J., et al., *A novel hybrid system for feature selection based on an improved gravitational search algorithm and k-NN method.* Applied Soft Computing, 2015. 31: p. 293-307.
[50] Xue, B., M. Zhang, and W.N. Browne, *Particle swarm optimisation for feature selection in classification: Novel initialisation and updating mechanisms.* Applied Soft Computing, 2014. 18: p. 261-276.
[51] Jaddi, N.S., J. Alvankarian, and S. Abdullah, *Kidney-inspired algorithm for optimization problems.* Communications in Nonlinear Science and Numerical Simulation, 2017. 42: p. 358-369.