Journal of Informatics Engineering (JUTIF) Vol. 6, No. 6, December 2025, Page. 5386-5400
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4728

Validation of Question Classification Using Support Vector Machine and Intraclass Correlation Coefficient Based on the Revised Bloom's Taxonomy

Lazuardy Syahrul Darfiansa*[1], Sza Sza Amulya Larasati[2]

[1]Informatics Telkom University, Indonesia
[2]Computer Science Brawijaya University, Indonesia

Email: [1]lazuardysyahrul@telkomuniversity.ac.id

Received : May 14, 2025; Revised : Jun 16, 2025; Accepted : Jun 16, 2025; Published : Dec 22, 2025

Abstract

The assessment process must be carried out accurately as it is a crucial aspect of identifying cognitive abilities in students. Cognitive ability identification needs to be done by providing exam questions that refer to the Revised Bloom's Taxonomy for difficulty-level classification to ensure students' understanding of what has been taught. The traditional manual classification process carried out by educators often requires significant time and is susceptible to subjective variability. The classification of questions from levels C1 to C6 based on the Revised Bloom's Taxonomy shows an imbalance in the data distribution for each level, leading to inaccurate classification results. The automatic classification technique using the SVM algorithm allows educators to quickly classify questions based on their difficulty levels. The automated classification technique needs to be validated to what extent the difficulty levels classified by the machine align with the perceptions of educators and students. This research will validate the results of question classification generated from the SVM algorithm, supplemented by the oversampling technique to address data imbalance. The validation method used is ICC. Applying the SMOTE oversampling technique to handle a class imbalance in the training data shows improvement, with an accuracy rate of 91% when using SMOTE compared to 83% without it. Results of the classification suitability test with the SVM algorithm by educators and students indicate a high level of agreement. The ICC Average Measures values are as follows: SVM classification is 0,979, assessment by non-science subject educators is 0,956, assessment by science subject educators is 0,991, assessment by non-science subject students is 0,982, and assessment by science subject students is 0,984. ICC testing consistently yields excellent results in non-science and science subjects, indicating that the assessments conducted by educators and students have a very high level of agreement.

Keywords : *ICC, Machine Learning, Question Classification, Revised Bloom Taxonomy, SVM*

This work is an open access article and licensed under a Creative Commons Attribution-Non Commercial 4.0 International License

1. INTRODUCTION

In the learning process, there is an assessment process, which is presented in the form of an exam. Exams are activities to measure competency achievement as a measure of student learning achievement. The assessment process must be carried out appropriately because it is essential in identifying students' cognitive abilities [1], [2], [3]. To ensure students understand what has been taught, it is necessary to understand cognitive abilities by providing exam questions that refer to the Revised Bloom's Taxonomy to classify difficulty levels. Bloom's Taxonomy is a conceptual framework for identifying thinking ability competencies from the lowest to the highest levels [4], [5], [6], [7]. The classification process for each question based on Bloom's Taxonomy can be carried out by identifying lexical and syntactic features [4], [5], [8]. Bloom's Taxonomy focuses on cognitive learning and knowledge. However, as time went by, the need emerged to refine the framework to make it more responsive to developments in education and technology. Revised Bloom's Taxonomy emerged as an attempt to modernize and expand Bloom's Taxonomy. This revision includes three main dimensions: Knowledge, Cognitive Processes, and Knowledge Products. The Knowledge Dimension includes the levels of remember, understanding,

and application. The Cognitive Process involves six stages, from remember to creation. Meanwhile, Knowledge Products involve concepts, procedures, and knowledge. Thus, the Revised Bloom's Taxonomy goes more in-depth in detailing the cognitive aspects and replaces some of the original terms, such as "Knowledge," which becomes "Remember," and "Synthesis," which becomes "Evaluation" shown in Figure. 1.

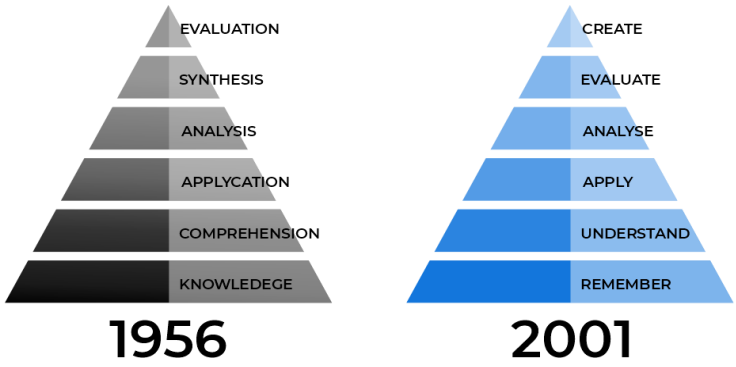

Figure 1. Cognitive Domain of Bloom Taxonomy

The traditional process of classifying exam questions as part of learning evaluation is generally done manually by educators. This manual classification requires educators to carefully read, interpret, and assign each question to the appropriate cognitive level based on BT, which is a highly subjective and labor-intensive task [9], [10], [11]. When dealing with large sets of questions such as in national exams or institutional assessments, this process becomes extremely time-consuming and inefficient. Moreover, this process is prone to inconsistency and subjective bias due to differences in perception among educators [12], [13], [14] as each may perceive the difficulty level of a question differently. This subjectivity can lead to inconsistencies in classifying questions that should actually belong to the same cognitive level. To overcome this issue, automatic question classification becomes a crucial solution to improve efficiency and ensure consistency in educational evaluations. Intraclass Correlation Coefficient (ICC) can support classifying question by measuring and validating the level of agreement between human raters. By adopting automation technology, such as Natural Language Processing (NLP) models, educators can classify exam questions more quickly and accurately based on Bloom's Taxonomy [15], [16], [17]. A widely used technique in NLP is Term Frequency-Inverse Document Frequency (TF-IDF) as a feature extractor. TF-IDF assigns a weight to each word based on the frequency of occurrence of the word in the whole corpus, thus being well able to filter out words that become specific features in a class [18], [19], [20]. Extracted features can be utilized as input into machine learning algorithms to perform automatic classification. By combining the advantages of the ICC approach, TF-IDF method, and machine learning-based classification techniques, the question classification process is possible in a more objective and efficient manner.

Several previous studies highlighted machine learning methods in text classification. Research by Callista et al. compared the classification results at levels C1 to C3 based on BT using Support Vector Machine (SVM) and Naive Bayes (NB). As a result, SVM gets a higher accuracy of 98% compared to NB of only 91%. A similar study by Ababneh [21] also compared the performance of NB, Random Forest (RF), KNN, SVM, and LR models in text classification. SVM got the highest accuracy of 82%, then LR with accuracy of 67.62%, RF of 58.5%, NB of 58%, and KNN of 57.75%. Another study by Luo [22] compared the performance of SVM, LR, and NB models. The highest accuracy was obtained in the SVM model of 86%, followed by LR with 81% accuracy, and NB by 24%. Research by Sabri

[23] also performs text classification by analyzing sentiment, detecting trends in customer feedback, detecting spam, labeling topics, and making comparisons. The classification models employed are SVM, DT, RF, KNN, and LR. SVM still provides the highest accuracy of 92.93%, followed by LR at 91.49%, KNN at 91.08%, RF at 90.94%, and DT at 80.02%. Further research by Mohammedid & Omar [24] also highlighted the effectiveness of using TF-IDF feature extraction techniques on SVM, LR, and KNN models. As a result, SVM outperformed the other two models with 89.7% accuracy followed by LR at 89.4% and KNN at 85.4%. In text classification tasks, imbalanced data distribution is a common challenge that can negatively impact model performance. A study by Mohasseb et al. [13] successfully addressed this issue by applying the Synthetic Minority Over-sampling Technique (SMOTE) to balance the amount of data between classes. The results showed that the accuracy of the model increased significantly, reaching 84.7% after the application of SMOTE, compared to the initial accuracy without SMOTE which was only 81.8%. Similar insights are also found in Callista's research, which states that the use of SMOTE can increase model accuracy up to 38.03%.

Existing literature on text classification tasks shows that SVM consistently outperforms the accuracy of other models such as KNN, DT, NB, RF, and LR. Most of these studies focus on machine learning algorithms that rely heavily on conventional feature extraction techniques such as TF-IDF, but the feature extraction process is not widely discussed. In the same time, although data distribution imbalance has been recognized as a common obstacle in classification, only a few studies have applied data balancing strategies through SMOTE. Classification processes specifically based on Revised BT are also rarely explored. Automatic classification techniques need upholding to validate the difficulty level of the machine classification according to educators' and learners' perceptions. All these research gaps will be resolved in this study. The classification process is specifically referenced through 6 classes in Revised BT by adapting SMOTE to balance the distribution of data in each class, performing feature extraction with TF-IDF, using SVM as a classifier that was proven excellent in text classification in general, until validating the classification results with human perception through ICC. Therefore, this research aims to validate an automatic question classification system based on Revised Bloom's Taxonomy using SVM, SMOTE for data balancing, and ICC for validation.

2. RELATED WORK

Research by Ababneh [21] aims to classify text in Arabic using a dataset of Arabic online news. The algorithms used to perform the classification include NB, RF, KNN, SVM, and LR. The results of this study show that the SVM algorithm produces 82% accuracy followed by LR 67.62%, RF 58.5%, NB 58%, and KNN 57.75%. The drawback is related to the relatively small number of categories in the Arabic dataset compared to the existing English dataset. The Arabic dataset only has seven categories as its maximum number. All these categories describe general topics that can be subdivided into sub-categories. The limited number of categories in this dataset can affect the performance of the classification model because it does not cover a wider diversity of topics such as those in the English dataset. The research conducted by Luo [22] aims to classify text and documents in English and compare the efficiency of several algorithms such as LR, NB, and SVM with a total of 1033 data. The results of the SVM research resulted in 86% accuracy, followed by LR 81%, and NB 24%. Despite the high accuracy, there are some drawbacks, namely that it does not provide an in-depth analysis of what factors cause the difference in performance between the learning models used. An analysis could help understand the reasons behind SVM's superiority over other models and provide valuable insights for future research. It does not specifically identify the weaknesses of each model used in the context of the dataset analyzed.

Research conducted by Mohasseb [13] aims to compare the use of optimization of classification data that is not balanced (oversampling) SMOTE on the NB algorithm. The accuracy results using SMOTE were 84.7% while those that did not use 81.8%. The drawback is that it does not provide a

detailed explanation of the evaluation method used to measure the performance of the proposed framework. Further information about the dataset used was not included, such as the number of unbalanced classes and the overall size of the dataset. This information is important to understand the characteristics of the data used in the study and its impact on the results. Research conducted Campos et al. [25] aims to determine the effect of word restrictions on text classification using the NB, DT, RF, and SVM algorithms, and compare the performance of each algorithm. The results showed that word restriction did not affect the classification results with SVM producing 93.04% accuracy, followed by DT 87.61%, NB 81.25%, and RF 80.74%. The drawback is that it does not provide enough detailed information about the methods and criteria used to select the proposed word restriction models. This lack of information makes it difficult for readers to assess the validity and relevance of the models used in the study. The word limitation models are not explained, which is very important to understand how each model works and its characteristics.

Research related to the classification of questions based on Bloom's Taxonomy levels C1 - C3 using SVM and NB algorithms. TF-IDF feature extraction technique was used to weight the words and oversampling technique (SMOTE) was used. Datasets that use oversampling techniques show higher classification results. The accuracy results using oversampling techniques, the SVM algorithm produces 98% and NB 91% while those that do not use oversampling techniques produce 77% and NB 71% Callista et al. [26]. Research by Mohammed and Omar [27] related to the classification of questions based on Bloom's Taxonomy using the TF-IDF feature extraction technique. The classification algorithms used are KNN, LR, and SVM. There are two datasets used containing 141 questions and 600 questions. The results of the first dataset containing 141 questions, the three classifiers achieved an average accuracy of 71.1%, 82.3%, and 83.7%. The second dataset of 600 questions achieved accuracies of 85.4%, 89.4%, and 89.7% respectively.

Research by Ismunarti et al., [28] looked for the most appropriate test statistics for reliability testing of instruments for aquatic chlorophyll-a concentrations and stated that the most appropriate test used was ICC. Water chlorophyll-a concentration data was obtained from two instruments, namely spectrophotometric methods from 14 observation stations in Semarang Bay, Central Java and remote sensing. The results obtained ICC value = 0.83 means that 83% of data variability is due to object diversity. ICC is more appropriate to test the reliability and validity of the instrument than the correlation coefficient, MAE and RMSE. Research by Zhao et al., [29] aims to assess how consistent the Chinese version of the Action Research Arm Test (C-ARAT) is when used by different raters (inter-rater) and by the same rater at different times (intra-rater) in patients recovering from their first stroke. The participants were inpatients receiving care at the Department of Rehabilitation Medicine, First Affiliated Hospital of Sun Yat-sen University. To measure reliability, the researchers used the Intra-class Correlation Coefficient (ICC). The results showed an ICC value of 0.998, indicating an extremely high level of consistency in how different raters assessed C-ARAT scores in these stroke patients.

3. METHOD

Proposed method to classify question based on Revised BT are shown in Figure 2. The stages have several parts as follows: (1) Dataset: Collecting question data needed to perform question classification. This dataset contains question data that has been labeled according to the level of the Revised BT. (2) Preprocessing: The text data in the dataset processed by performing steps such as special character removal, text conversion to lowercase, and removal of stopwords (common words that do not contribute significantly to the analysis). (3) Feature Extraction: The preprocessed text data will be converted into a number vector representation using TF-IDF Weighting Scheme method. (4) Split Data: The dataset are divided into two parts as training data and testing data with a ratio of 80:20. The training data are used to train the model, while the testing data will be used to test the performance of the trained model. (5) Oversampling: The data are imbalance, oversampling technique utilized using SMOTE with default parameters. (6) Classification: The classification process is done using SVM. SVM is classify the data based on the features generated from the previous stages. (7) Evaluation: Evaluation of SVM

performance, using evaluation metrics such as confusion matrix, accuracy, precision, recall, and F1-Score to measure how well the SVM model performs problem classification. (8) Intraclass Correlation Coefficient: Evaluates the extent of agreement between SVM classification results, educator assessments, and student assessments.

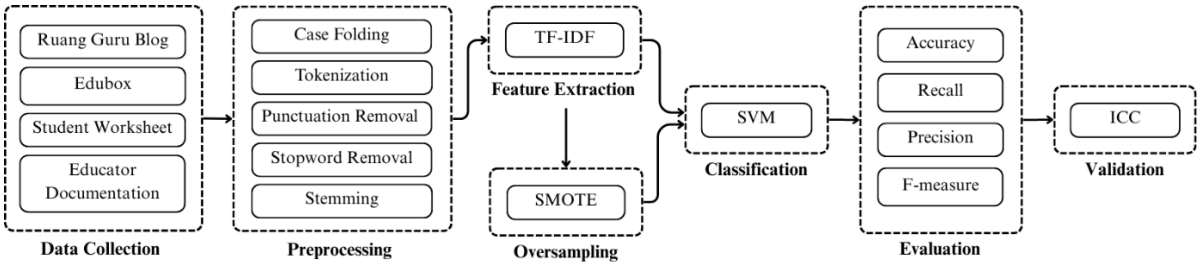

Figure 2. Methodology

3.1. Data Collecting

This research uses a dataset consisting of 1,322 Indonesian questions that were primarily collected from various online and offline educational sources shown in Figure 3. These sources include the Ruangguru Blog, Edubox, student worksheet, and educator documentation. The questions span a variety of subjects and grade levels to provide diverse cognitive challenges. The distribution of the collected data is imbalance and presented in Figure 1. Each question was then manually labeled by two education experts according to the cognitive levels defined in the Revised BT of six categories: C1 (Remember), C2 (Understand), C3 (Apply), C4 (Analyze), C5 (Evaluate), and C6 (Create). The question labeling process uses the following criteria:

a. C1 if the question asks students to remember facts, definitions, or basic information.
b. C2 if the question requires students to explain, interpret, or summarize the concept or information provided.
c. C3 if the question asks students to apply the knowledge or concepts they have learned to a new situation or problem.
d. C4 if the question requires students to describe, dissect, or arrange elements of concepts or information in a more complex relationship.
e. C5 if the question asks students to evaluate, decide, or determine the relative value of various options or arguments.
f. C6 if the question asks students to design, create, or develop something new based on the concepts or knowledge they have been given.

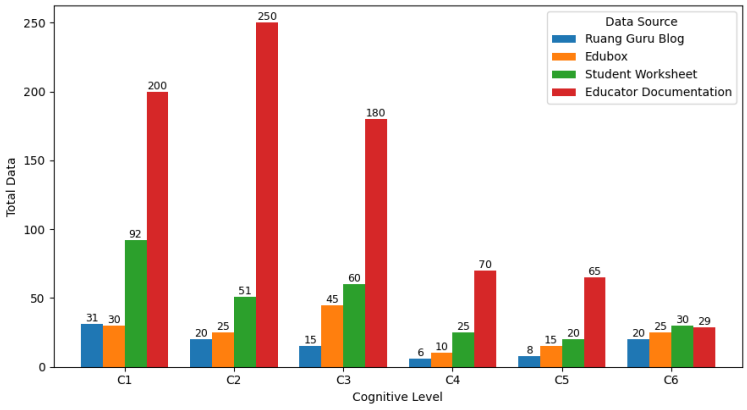

Figure 3. Data distribution

3.2. Preprocessing

All labeled question texts undergo a preprocessing phase to clean, standardize, and ensure data concistency before being used as training and testing data. Firstly, all characters are converted to lowercase to eliminate inconsistencies caused by letter casing on case folding step, then each question are breaking down into individual words or tokens on tokenization step. These tokens are passed through a punctuation removal step to eliminate symbols, marks, characters, and number that do not contribute meaningful information. Common Indonesian words such as "*yang*", "*dan*", and "*di*" that do not affect the semantic interpretation are filter out on stopword removal step using NLTK Indonesia library from Sastrawi. On the last step, stemming is implemented to maintain the root word by removing the affixes in each word.

3.3. Feature Extraction

The feature extraction process is leveraging TF-IDF. This method measures how often a word appears in a document (TF) and the extent of how infrequently it appears in the entire set of documents or corpus (IDF). The combination of these two components generates a numerical weight that emphasizes the importance of a word in a single document relative to the entire corpus. In BT, each cognitive level from C1 to C6 has different linguistic features, such as operational verbs or specific sentence structures that reflect the intended thought activity. For example, words like "*sebutkan*" and "*definisikan*" often appear in C1 questions, while words like "analyze", "design", or "evaluate" are more common in higher-level questions like C4, C5, and C6. TF-IDF gives more weight to the key words that reflect each cognitive level, thus aiding in the automatic classification process based on the desired level of thinking.

3.4. Oversampling Testing

The dataset employed has an unbalanced class distribution. Class C1 has the largest amount of data with 353 data being the dominant class in model training. This dominance can affect the training process because the model tends to recognize patterns in class C1 more than other classes. As a result, the evaluation results on class C1 can be extremely good, while the performance on other classes becomes lower. To overcome this imbalance, an oversampling process is performed on classes C2 to C6 using the SMOTE method from the Imbalanced-Learn library. SMOTE generates synthetic data for minority classes by utilizing the KNN concept (Mohasseb et al., 2018). The SMOTE process works by selecting a sample from the minority class, finding the k nearest neighbors based on distance then creating new data by randomly combining attributes from the initial sample and its neighbors. The SMOTE parameters applied are the default parameters as shown in Table 1.

Table 1. SMOTE Parameters

Parameter	Value
sampling_strategy	auto
random_state	none
k_neighbors	5
n_jobs	none

3.5. Classification

The SVM algorithm is used to build a model that is able to classify questions so that they fall into the appropriate criteria based on the training process using train data and evaluated by predicting using test data. This research uses a special SVM class for classification, namely SVM which comes from the Scikit-Learn library. The parameters used are the default parameters as in Table 2. The question data was divided separately into 80% data as training data and the remaining 20% as testing data.

Table 2. SVM Parameters

Model	Value
C	1.0
Kernel	RBF
Gamma	3
Coef0	Scale
Shrinking	True
Probability	False
Tol	1e-3
Cache_size	200
Class_weight	None
Verbose	False
Max_iter	-1
decision_function_shape	ovr
break_ties	False
Random_state	None

3.6. Evaluation

A confusion matrix is a handy way to check how well a classification model is doing [23]. It shows not just which predictions were right or wrong, but also what kind of mistakes the model made. It includes four main parts: true positives (TP), when the model correctly predicts something as positive; true negatives (TN), when it correctly predicts something as negative; false positives (FP), when it wrongly says something is positive; and false negatives (FN), when it misses something that should be positive. By looking at these, we can calculate performance scores like accuracy, recall, precision, and F-measure. These scores help us understand where the model is doing well and where it might need improvement. If the recall or F-measure is low, it might mean that the dataset isn't balanced and some categories don't have enough examples.

3.7. Validation

ICC is a statistic used to measure the reliability of a measurement on one continuous variable carried out using two or more different measuring instruments [30], [31]. Measurements are carried out with two measuring instruments on discrete-sized variables, so the statistical test generally used is Cohen's kappa coefficient of agreement. ICC is the ratio between the variance between groups and the total variance. The total variance comes from 3 sources: (1) actual values, (2) assessor (SVM classification, educator), and (3) random error (residual error). Observer variation is assumed to be random, so the ICC formula is following Equation (1)

$$ICC = \frac{\sigma \frac{2}{s}}{\sigma \frac{2}{s} + \sigma \frac{2}{o} + \sigma \frac{2}{e}} \quad (1)$$

The ICC value describes the variance ratio between measurement objects and the total variance. The range of ICC values is between zero and one ($0 \leq ICC \leq 1$). An ICC value close to one indicates that the reliability of the measuring instrument is approaching perfection, meaning that most of the data variation is attributed to differences between the objects being measured rather than inconsistencies among the measuring instruments. ICC values close to zero or low can result in inconsistencies in measuring instruments, instability of the object being measured, and unsupported measurement situations [32]. A low ICC value can indicate a lack of consistency between measuring devices or the

presence of external factors that influence measurement results, such as differences in measurement conditions. ICC provides an overview of how well an instrument can be relied upon to measure a particular variable. According to Giuseppe Via San Lorenzo [32], ICC statistical criteria are shown in Table 3.

Table 3. ICC statistics criteria

ICC	Criteria
ICC < 0,50	Poor
0,50 ≤ ICC < 0,75	Moderate
0,75 ≤ ICC < 0,90	Good
ICC > 0,90	Excellent

ICC measurements were conducted through 80 assessments from students and 20 educators in Mathematics subjects. This process involves giving students and educators a questionnaire sheet consisting of 24 questions divided from difficulty levels C1 to C6, with four questions for each difficulty level. Details of the sample questions given are shown in Table 4. Through this questionnaire, validation involved students' and educators' perceptions of various aspects of the mathematics subject. The ICC results from this assessment provide an overview of the extent of conformity between actual scores, classification results, student, and educator.

Table 4. Question sample

No	Question	Category					
		C1	C2	C3	C4	C5	C6
1	**C1 - Remembering** *Sebutkan dua jenis bilangan bulat.*						
2	**C2 - Understanding** *Bagaimana cara menghitung rata-rata dari sekumpulan angka?*						
3	**C3 - Applying** *Hitunglah keliling segitiga siku-siku dengan panjang sisi-sisi 3 cm, 4 cm, dan 5 cm.*						
4	**C4 - Analyzing** *Jika A = {1, 2, 3, 4, 5} dan B = {3, 4, 5, 6, 7}, tentukan irisan (intersection) A dan B.*						
5	**C5 - Evaluating** *Sebuah perusahaan menawarkan dua rencana berlangganan internet: A seharga Rp 200,000/bulan dengan kecepatan 50 Mbps dan B seharga Rp 250,000/bulan dengan kecepatan 100 Mbps. Manakah rencana yang lebih ekonomis berdasarkan harga per Mbps?*						
6	**C6 - Creating** *Untuk jumlah 6036 suku pertama deret geometri adalah 1141 dan jumlah 4024 suku pertamanya sama degan 780, maka jumlah 2012 suku pertamanya adalah*						

4. RESULT AND DISCUSSION

This section presents the results obtained from the experiments and provides a comprehensive discussion of the findings. The performance of the classification model is evaluated based on various

metrics, and the impact of different hyperparameter settings is analyzed. The discussion highlights the strengths and potential limitations of the model in addressing the question classification task, offering insights into its practical implications and areas for future improvement

4.1. Performance of SVM

The evaluation results of the model performance in each class show an unevenness between classes as shown in Table 5. Class C1 shows a strong and consistent performance with a precision value of 0.86, recall of 0.87, and F1-score of 0.86. It implies that the model is well generalized to the patterns in this class. This high performance is strongly suspected to be the larger amount of data in class C1 compared to the other classes, affording the model more learning opportunities. Classes C2 and C3 also still perform quite well with F1-score of 0.80 and 0.90 respectively. Particularly in class C3, the recall reaches the highest value of 0.92, implying that the model is proficient in recognizing most of the instances from this class. However, based on the confusion matrix, there are still misclassifications between classes C1, C2, and C3, possibly due to the overlap of language or semantic structure among the initial cognitive levels such as remembering, understanding, and applying.

In contrast, the model exhibited lower performance in handling higher cognitive level classes such as C4, C5, and C6. The confusion matrix for this unbalanced data test is shown in Figure 5 (a). Recall on the C4 class was only 0.59, with 9 instances classified as C2. This phenomenon shows that questions that demand analytical skills are often mistaken for basic understanding. Class C6 also had a low recall of 0.52 despite perfect precision of 1.00, which suggests that the model only predicted this class when it was very confident, but often failed to recognize it with only 11 out of 21 instances correctly classified. Class C5 has a high recall of 0.91, but a precision of only 0.71, which indicates that this class often receives incorrect predictions from classes C4 and C6. Although the overall accuracy of the model reached 83%, this number disguises the performance gap between classes. This finding shows that the model is still biased towards the majority class or the class with more easily distinguishable language characteristics and is not sensitive enough to recognize the characteristics of questions in the minority class or high cognitive level.

Table 5. Classification report of SVM before oversampling

Classes	Precision	Recall	F1-Score	Accuracy
C1	0.86	0.87	0.86	83%
C2	0.76	0.85	0.80	
C3	0.89	0.92	0.90	
C4	0.87	0.59	0.70	
C5	0.71	0.91	0.80	
C6	1.00	0.52	0.69	

To overcome the discrepancies in inter-class classification results, SMOTE is undertaken to equilibrate the data distribution in each class. The data distribution before and after SMOTE is shown in Table 6. After oversampling using SMOTE, the performance of the model improved in almost all classes, especially in minority classes. The overall accuracy increased from 83% to 91% after oversampling. The comparison of matrix evaluation before and after oversampling is shown in Figure 4. This improvement shows that a more equal distribution of data can help the model to learn more fairly and thoroughly for all classes.

Journal of Informatics Engineering (JUTIF) Vol. 6, No. 6, December 2025, Page. 5386-5400
P-ISSN: 2723-3863 https://jutif.if.unsoed.ac.id
E-ISSN: 2723-3871 DOI: https://doi.org/10.52436/1.jutif.2025.6.6.4728

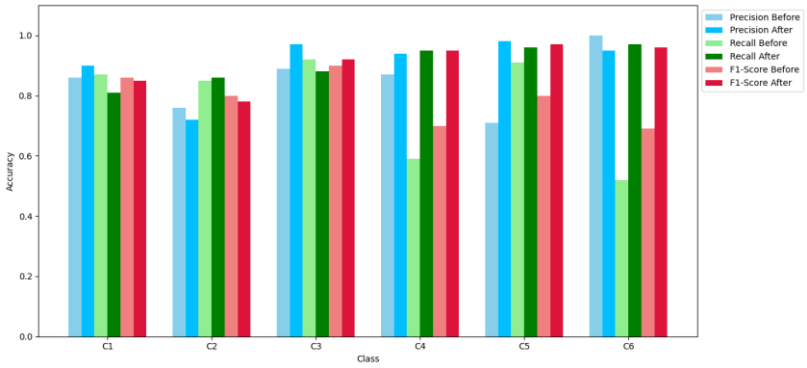

Figure 4. Classification reporth

In classifications with imbalanced distribution between classes, recall is considered as evaluation metric because errors in detecting the minority class (FN) have more impact on the overall performance of the system, especially when each class is considered to be of equal importance. Recall is used to find as many instances of the minority class as possible that were previously misclassified due to data imbalance. According to the test results in Table 5 and confusion matrix in Figure 5 (b) after oversampling, the recall value increases greatly in the previously minority classes, specifically classes C4, C5, and C6. In class C6, recall value increases from 0.52 to 0.97. Out of 62 instances in class C6, 60 instances were classified correctly. Similarly, class C4 which initially had a recall of 0.59 increased to 0.95, and class C5 from 0.91 to 0.96. The model correctly classified 80 out of 84 instances in class C4, and 80 out of 83 instances in class C5. Not only has the performance of the minor classes improved, but also the major classes such as C2 and C3 still have a stable high performance with recall values of 0.86 and 0.88. The improvement shows that the model is now better able to recognize patterns from minor classes that were previously overlooked.

Table 6. Oversampling distribution

Classes	Before	After
C1	353	353
C2	346	353
C3	300	353
C4	111	353
C5	108	353
C6	104	353
Total	**1,322**	**2,118**

Table 7. Classification report of SVM after oversampling

Classes	Precision	Recall	F1-Score	Accuracy
C1	0.90	0.81	0.85	91%
C2	0.72	0.86	0.78	
C3	0.97	0.88	0.92	
C4	0.94	0.95	0.95	
C5	0.98	0.96	0.97	
C6	0.95	0.97	0.96	

The model successfully classified 54 instances correctly out of a total of 63 instances in class C2, and 56 instances out of 64 instances in class C3. Compared to C4, C5, and C6, misclassification

occurred more in classes C1, C2, and C3. Competency C1 emphasizes recalling and reciting information, C2 involves understanding the meaning and interpretation of information, and C3 involves applying information in new situations. The thought processes in these three classes are interrelated and tend to use similar representations of information, so the extracted features by the model are likely to overlap and have ambiguous inter-class boundaries, especially in the data synthesized using SMOTE based on interpolation between nearest neighbors. In class C1, 9 instances were misclassified as C2 out of 68 actual instances of C1. Similarly, the C2 class misclassified 5 instances as C1 and 6 instances of C3 instead of C2.

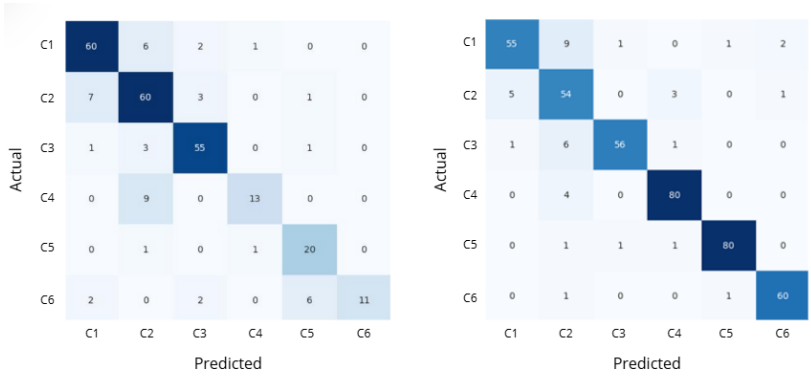

Figure 5. Confusion matrix before (a) and after (b) the implementation of SMOTE

4.2. Performance of ICC

ICC testing involves assessments from educators and students to measure the level of conformity or agreement between assessors. Values range between 0 and 1, with 1 indicating perfect agreement and 0 indicating low agreement. Actual Results are the results of question classification carried out by educators who experts in classifying questions are based on the Revised BT. SVM results are the results of question classification using this model. The results of the ICC test in Table 6 for educators show variations in reliability between pairs of assessors. Some pairs showed a high level of agreement; for example, the agreement value of assessors 1 with the actual value and SVM got a value of 0.94 and 0.955, while the low agreement value of assessors 5 with the actual value and SVM got a value of 0.340 and 0.409. Table 7 provides detailed statistics for the ICC generated from Table 8. ICC Single Measures of 0.644 indicates moderate reliability for one measurement, while ICC Average Measures of 0.956 indicates high reliability for several measurements. Based on [24], the ICC statistical criteria shown in Table 9 show that the ICC Average Measures value of 0.956 agrees that the assessment of subject educators is in the excellent category.

Table 8. Sample results of the ICC test for educators

	Actual Value	SVM	Assessor1	Assessor2	Assessor3	Assessor4	Assessor5
Actual Value	1.000	0.964	0.986	0.899	0.928	0.964	0.964
SVM	0.964	1.000	0.948	0.850	0.894	0.924	0.941

Table 9. Detailed statistics result ICC educators

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	0.746[a]	0.631	0.855	76.147	23	483	0.000
Average Measures	0.985[c]	0.974	0.992	76.147	23	483	0.000

Testing on students in science subjects involved 80 students in Table 10, representing five samples of ICC testing results. The ICC test results on a sample of students mostly scored below 0.5 and were included in the poor category. The agreement value between Actual, SVM, and Assessors 1, 3, and 4 obtained a value between 0.75 and 0.90, which is included in the moderate category, indicating that the level of agreement with the classification results is high. Table 11 provides an overall summary of assessors regarding ICC results. ICC Single Measures is 0.580, indicating moderate reliability for one measurement, while ICC Average Measures is 0.991, indicating a very high level of agreement for several measurements that are in the excellent category. The limited 95% confidence interval, between 0.985 and 0.996, shows that the assessments made by several assessors have a very high level of consistency and uniformity

Table 10. Sample results of the ICC test for students

	Actual Value	SVM	Assessor1	Assessor2	Assessor3	Assessor4	Assessor5
Actual Value	1.000	0.964	-0.079	0.645	0.502	0.623	0.623
SVM	0.964	1.000	-0.042	0.653	0.506	0.616	0.616

Table 11. Detailed statistics result ICC students

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	0.580[a]	0.452	0.732	116.739	23	1863	0.000
Average Measures	0.991[c]	0.985	0.996	116.739	23	1863	0.000

The present research highlights the contribution to educational technology by emphasizing the effectiveness of machine learning techniques using SVM with SMOTE method in classifying questions based on the Revised Bloom's Taxonomy. The improved accuracy with SMOTE and the model's perceptual agreement with humans reveal how the proposed model provides reliable and objective support in evaluating the difficulty level of questions. Nonetheless, this study is limited to the application of SVM (using default hyperparameters) with a maximum accuracy of 91%. The feature extraction process is also limited to TF-IDF that may not fully capture the semantic depth in natural language.

5. CONCLUSION

Revised BT is a benchmark in classifying questions into 6 cognitive levels. In real cases, the distribution of data is generally unbalanced so it needs specific handling to improve the predictive ability of the model. The application of the SMOTE oversampling technique that was implemented increased the accuracy from 83% to 91% using the SVM classifier. The results of testing the suitability of SVM classifications with educators and students show a high level of agreement. The ICC Average Measures value is SVM classification of 0.979, assessment by non-science subject educators 0.956, assessment by science subject educators 0.991, assessment by non-science subject students 0.982, and assessment by science subject students 0.984. This high level of agreement suggests that SVM provides consistent classification results and can be accepted by both assessors, both educators and students, in assessing the difficulty of the questions. This research provides a meaningful contribution to educational technology by demonstrating the machine learning application, especially SVM combined with data balancing techniques, can support the question classification process in an objective and consistently measurable manner. To improve the understandability of complex semantic structures in queries, future research could adapt more sophisticated models such as CNN, LSTM, or transformer-based models. These approaches offer different feature extraction processes that produce variations in classification performance, so further exploration could open up opportunities to improve model accuracy.

CONFLICT OF INTEREST

The authors declares that there is no conflict of interest between the authors or with research object in this paper.

ACKNOWLEDGEMENT

Acknowledgement is only addressed to funders or donors and object of research. Acknowledgement can also be expressed to those who helped carry out the research.

REFERENCES

[1] E. Ita, "Manajemen Pembelajaran Pendidikan Anak Usia Dini di TK Rutosoro Kecamatan Golewa Kabupaten Ngada Flores Nusa Tenggara Timur," *J. Dimensi Pendidik. Dan Pembelajaran*, vol. 6, Jun. 2018, doi: 10.24269/dpp.v6i1.889.

[2] E. R. Setyaningsih and I. Listiowarni, "Categorization of Exam Questions based on Bloom Taxonomy using Naïve Bayes and Laplace Smoothing," in *2021 3rd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, 2021, pp. 330–333. doi: 10.1109/EIConCIT50028.2021.9431862.

[3] A. Sangodiah, T. J. San, Y. T. Fui, L. E. Heng, R. K. Ayyasamy, and N. B. A. Jalil, "Identifying Optimal Baseline Variant of Unsupervised Term Weighting in Question Classification Based on Bloom Taxonomy," *Mendel*, vol. 28, no. 1, pp. 8–22, 2022, doi: 10.13164/mendel.2022.1.008.

[4] I. Magdalena, N. F. Islami, E. A. Rasid, and N. T. Diasty, "Tiga Ranah Taksonomi Bloom dalam Pendidikan," *J. Edukasi Dan Sains*, vol. 2, no. 1, pp. 132-`139, Jun. 2020, doi: 10.36088/EDISI.V2I1.822.

[5] E. Subiyantoro, A. Ashari, and Suprapto, "Cognitive Classification Based on Revised Bloom's Taxonomy Using Learning Vector Quantization," in *CENIM 2020 - Proceeding: International Conference on Computer Engineering, Network, and Intelligent Multimedia 2020*, Institute of Electrical and Electronics Engineers Inc., Nov. 2020, pp. 349–353. doi: 10.1109/CENIM51130.2020.9297879.

[6] I. I. Sinan, V. Nwoacha, J. Degila, and S. A. Onashoga, "A Comparison of Data-Driven and Data-Centric Architectures using E-Learning Solutions," *Proceedings - International Conference Advancement in Data Science, E-Learning and Information Systems, ICADEIS 2022*, 2022, doi: 10.1109/ICADEIS56544.2022.10037358.

[7] K. Makhlouf, L. Amouri, N. Chaabane, and N. El-Haggar, "Exam Questions Classification Based

on Bloom's Taxonomy: Approaches and Techniques," in *2020 2nd International Conference on Computer and Information Sciences, ICCIS 2020*, Institute of Electrical and Electronics Engineers Inc., Oct. 2020. doi: 10.1109/ICCIS49240.2020.9257698.

[8] H. Shi *et al.*, "Educational management in Critical Thinking Training Based on Bloom's Taxonomy and SOLO Taxonomy," *Proc. - 2020 Int. Conf. Inf. Sci. Educ. ICISE-IE 2020*, pp. 518–521, Dec. 2020, doi: 10.1109/ICISE51755.2020.00116.

[9] H. Sharma, R. Mathur, T. Chintala, S. Dhanalakshmi, and R. Senthil, "An effective deep learning pipeline for improved question classification into bloom's taxonomy's domains," *Educ Inf Technol (Dordr)*, vol. 28, no. 5, pp. 5105–5145, May 2023, doi: 10.1007/s10639-022-11356-2.

[10] M. O. Gani, R. K. Ayyasamy, A. Sangodiah, and Y. T. Fui, "Bloom's Taxonomy-based exam question classification: The outcome of CNN and optimal pre-trained word embedding technique," *Educ Inf Technol (Dordr)*, vol. 28, no. 12, pp. 15893–15914, Dec. 2023, doi: 10.1007/s10639-023-11842-1.

[11] L. S. Darfiansa and F. A. Bachtiar, "Comparative Analysis of Term Weighting Methods for Question Classification in Bloom Taxonomy Using Machine Learning Approach," in *2023 IEEE International Conference on Computing (ICOCO)*, IEEE, Oct. 2023, pp. 259–264. doi: 10.1109/ICOCO59262.2023.10397821.

[12] S. Basuki, Z. Rakhmawati, and G. W. Wicaksono, "Klasifikasi Kalimat Tanya Berdasarkan Taksonomi Bloom Menggunakan Support Vector Machine," *J. Repos.*, vol. 2, no. 4, pp. 427–436, 2020, doi: https://doi.org/10.22219/repositor.v2i4.30516.

[13] A. Mohasseb, M. Bader-El-Den, M. Cocea, and H. A. N. Liu, "Improving Imbalanced Question Classification Using Structured Smote Based Approach," in *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*, 2018, pp. 593–597. doi: 10.1109/ICMLC.2018.8527028.

[14] M. Ifham, K. Banujan, B. T. G. S. Kumara, and P. M. A. K. Wijeratne, "Automatic Classification of Questions based on Bloom's Taxonomy using Artificial Neural Network," *2022 International Conference on Decision Aid Sciences and Applications, DASA 2022*, pp. 311–315, 2022, doi: 10.1109/DASA54658.2022.9765190.

[15] N. Barari, M. RezaeiZadeh, A. Khorasani, and F. Alami, "Designing and validating educational standards for E-teaching in virtual learning environments (VLEs), based on revised Bloom's taxonomy," *Interactive Learning Environments*, vol. 30, no. 9, pp. 1640–1652, 2022, doi: 10.1080/10494820.2020.1739078.

[16] A. A. Yahya, Z. Toukal, and A. Osman, "Bloom's Taxonomy–Based Classification for Item Bank Questions Using Support Vector Machines," *Stud. Comput. Intell.*, vol. 431, pp. 135–140, 2012, doi: 10.1007/978-3-642-30732-4_17.

[17] K. Jayakodi, M. Bandara, and I. Perera, "An automatic classifier for exam questions in Engineering: A process for Bloom's taxonomy," in *2015 IEEE International Conference on Teaching, Assessment, and Learning for Engineering (TALE)*, 2015, pp. 195–202. doi: 10.1109/TALE.2015.7386043.

[18] L. S. Darfiansa, F. Azzuri, F. A. Bachtiar, and D. E. Ratnawati, "Comparative Analysis of Deep Learning and Machine Learning Techniques for Question Classification in Bloom's Taxonomy," *2023 1st International Conference on Advanced Engineering and Technologies, ICONNIC 2023 - Proceeding*, pp. 103–108, 2023, doi: 10.1109/ICONNIC59854.2023.10467502.

[19] M. Amien, "Sejarah dan Perkembangan Teknik Natural Language Processing (NLP) Bahasa Indonesia: Tinjauan tentang sejarah, perkembangan teknologi, dan aplikasi NLP dalam bahasa Indonesia," *ELANG J. Interdiscip. Res.*, vol. 1, no. 01, pp. 99–105, Aug. 2023, doi: 10.32664/ELANG.V1I01.

[20] Hasmawati, A. Romadhony, and R. Abdurohman, "Primary and High School Question Classification based on Bloom's Taxonomy," in *2022 10th International Conference on Information and Communication Technology, ICoICT 2022*, Institute of Electrical and Electronics Engineers Inc., 2022, pp. 234–239. doi: 10.1109/ICoICT55009.2022.9914842.

[21] A. H. Ababneh, "Investigating the relevance of Arabic text classification datasets based on supervised learning," *Journal of Electronic Science and Technology*, vol. 20, no. 2, p. 100160, Jun. 2022, doi: 10.1016/J.JNLEST.2022.100160.

[22] X. Luo, "Efficient English text classification using selected Machine Learning Techniques," *Alexandria Engineering Journal*, vol. 60, no. 3, pp. 3401–3409, Jun. 2021, doi: 10.1016/J.AEJ.2021.02.009.

[23] T. Sabri, O. El Beggar, and M. Kissi, "Comparative study of Arabic text classification using feature vectorization methods," *Procedia Comput Sci*, vol. 198, pp. 269–275, 2022, doi: https://doi.org/10.1016/j.procs.2021.12.239.

[24] M. Mohammed and N. Omar, "Question Classification Based on Bloom's Taxonomy Using Enhanced TF-IDF," *Int. J. Adv. Sci. Eng. Inf. Technol.*, vol. 8, no. 4–2, pp. 1679–1685, Sep. 2018, doi: 10.18517/IJASEIT.8.4-2.6835.

[25] D. Campos, R. R. Silva, and J. Bernardino, "Text mining in hotel reviews: Impact of words restriction in text classification," in *IC3K 2019 - Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, SciTePress, 2019, pp. 442–449. doi: 10.5220/0008346904420449.

[26] A. S. Callista, O. N. Pratiwi, and E. Sutoyo, "Questions Classification Based on Revised Bloom's Taxonomy Cognitive Level using Naive Bayes and Support Vector Machine," in *Proceedings - 2021 4th International Conference on Computer and Informatics Engineering: IT-Based Digital Industrial Innovation for the Welfare of Society, IC2IE 2021*, Institute of Electrical and Electronics Engineers Inc., 2021, pp. 260–265. doi: 10.1109/IC2IE53219.2021.9649187.

[27] M. Mohammedid and N. Omar, "Question classification based on Bloom's taxonomy cognitive domain using modified TF-IDF and word2vec," *PLoS One*, vol. 15, no. 3, 2020, doi: 10.1371/journal.pone.0230442.

[28] D. H. Ismunarti, M. Zainuri, D. N. Sugianto, and S. W. Saputra, "Pengujian Reliabilitas Instrumen Terhadap Variabel Kontinu Untuk Pengukuran Konsentrasi Klorofil- A Perairan," *Buletin Oseanografi Marina*, vol. 9, no. 1, pp. 1–8, Apr. 2020, doi: 10.14710/buloma.v9i1.23924.

[29] J. L. Zhao *et al.*, "Inter-rater and Intra-rater Reliability of the Chinese Version of the Action Research Arm Test in People With Stroke," *Front Neurol*, vol. 10, no. MAY, 2019, doi: 10.3389/FNEUR.2019.00540.

[30] T. K. Koo and M. Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," *J Chiropr Med*, vol. 15, no. 2, pp. 155–163, Jun. 2016, doi: 10.1016/J.JCM.2016.02.012.

[31] V. S. Senthil Kumar and S. Shahraz, "Intraclass correlation for reliability assessment: the introduction of a validated program in SAS (ICC6)," *Health Serv Outcomes Res Methodol*, pp. 1–13, Mar. 2023, doi: 10.1007/S10742-023-00299-X/TABLES/11.

[32] G. Perinetti, "StaTips Part IV: Selection, interpretation and reporting of the intraclass correlation coefficient," *South Eur. J. Orthod. Dentofac. Res. SEJODR*, vol. 5, no. 1, pp. 3–5, May 2018, doi: 10.5937/SEJODR5-17434.