SINTECHCOM: Science, Technology and Communication Journal
Vol. 6, No. 1, October 2025, pp. 33-44, DOI: 10.59190/stc.v6i1.337

A classification of Quran translations using K-nearest neighbors, support vector machine and random forest method

Nur Delifah*, Nazruddin Safaat Harahap, Surya Agustian, Muhammad Irsyad, Iwan Iskandar
Department of Informatics Engineering, UIN Sultan Syarif Kasim Riau, Pekanbaru 28293, Indonesia

ABSTRACT

A Classification of Quranic verses based on topics is one of the efforts to facilitate understanding and searching for information in the holy book, especially for non-Arabic readers. This study aims to test and compare the performance of three text classification methods, namely K-nearest neighbors (KNN), support vector machine (SVM), and random forest (RF), in grouping translated Quranic verses into 15 topic classes, such as Islamic arkanul, faith, the Quran, science and its branches, charity, da'wah, jihad, human and social relations, and others. The dataset used is the English translation of the Quran with full preprocessing and an 80:20 data split for training and testing. The evaluation was carried out using accuracy, precision, recall, and F1-score metrics. The results show that RF achieved the best performance with an average F1-score of 58.48% and testing accuracy of 90.81%. KNN followed with an F1-score of 54.07% and the highest testing accuracy of 92.05%, while SVM produced the lowest F1-score at 50.76% and accuracy of 88.20%. The RF demonstrates a more balanced ability in recognizing all classes, KNN excels in overall accuracy, and SVM performs less optimally in this classification task. This research is expected to serve as a foundation for developing a more intelligent and contextual topic-based verse classification system.

ARTICLE INFO

Article history:

Received Oct 3, 2025
Revised Oct 26, 2025
Accepted Oct 27, 2025

Keywords:

Classification
K-Nearest Neighbors
Quran Translation
Random Forest
Support Vector Machine

* **Corresponding Author**
E-mail address: 12250123957@students.uin-suska.ac.id

1. INTRODUCTION

The Qur'an is the main source of teachings in Islam which is believed by Muslims to be the true revelation, this holy book contains the words of Allah which were conveyed by the angel Gabriel to the Prophet Muhammad as the messenger of Allah in stages [1]. The purpose of the Qur'an is to be a guide for the lives of Muslims, helping them achieve happiness in this world and the hereafter [2]. The Qur'an consists of 114 parts called "Surahs" and is divided into 30 parts and consists of 6,236 verses called "ayah". Studying the Quran is the main obligation of every Muslim [3]. Learning and understanding the translation of the Qur'an is not easy, one way that can be done is to classify the translation of the verses of the Qur'an into existing topics.

Each verse in the Qur'an has a deep meaning and often covers a variety of themes at once. Based on the results of the study [4] , the content of the verses of the Qur'an can be categorized into 15 main topics, namely Arkanul Islam, Iman, Al-Qur'an, Science, Amal, Dakwah, Jihad, Human and Community relations, Akhlak, Rules relating to property, Matters relating to the law, Country and Society, Agriculture and Trade, History and Story, and Religions. The study used the Naïve Bayes multinomial method and succeeded in obtaining a hamming loss 0.1247.

To make it easier to understand and recognize the topics in the Qur'an, a system is needed that is able to group and identify verses into certain categories, known as the classification process. Text classification is one of the most important tasks in the field of natural language processing (NLP), where text is classified into predefined classes. There are various approaches in the classification of texts, but in this study the K-Nearest Neighbor (KNN), Support Vector Machine (SVM) and Random Forest (RF) methods were used. The KNN algorithm is known as an effective method for text classification tasks due to its simplicity, its ability to handle large amounts of data, as well as its ease of application [5]. Support Vector Machine (SVM) has the advantage of being able to handle high-dimensional data without relying on the number of attributes. In addition to being popular, SVM is also computationally efficient and effective in overcoming doubts in the learning process [6]. The classification process in Random Forest is carried out by combining a number of decision trees and training them using available sample data. This algorithm works by separating attributes from classes to generate predictions against new data that has not been recognized before. The use of the Random Forest algorithm gives an advantage in terms of flexibility and consistency when classifying data with more than one label [7].

KNN has been used in various studies for classifying Quranic verses and other Arabic texts. It generally performs well but is often outperformed by SVM in terms of accuracy. For instance, KNN achieved a recognition accuracy of 97.05% on the KUFIC dataset and 97.80% on the AHDB dataset [5]. However, in another study, KNN achieved an accuracy of 68% in classifying Indonesian translations of Quranic verses [8]. KNN has been applied to classify Quranic verses into categories such as "Shahadah" and "Pray" with more than 70% accuracy [9], and into categories like faith, worship, and etiquettes with above 80% accuracy [10].

SVM has been used to classify Quranic verses into themes and categories, achieving high accuracy scores. For example, it was used to classify verses into themes like faith, worship, and etiquettes with high accuracy [10]. Additionally, SVM was found to be effective in addressing the interrelationship between Quran and Hadith texts [11].

Random Forest is another robust classifier used in text classification tasks. It has been shown to improve classification performance when combined with feature extraction techniques like SIFT and SURF [12]. However, specific performance metrics for Quranic text classification using RF are not detailed in the abstracts. RF has been applied to classify ancient Arabic manuscripts and has shown promising results when synthetic features are introduced to enhance classification performance [12].

This study aims to compare the performance of the three methods in classifying the verses of the Qur'an into 15 classes, namely Islamic arkanul, faith, the Qur'an, science and its branches, charity, da'wah, jihad, human and social relations, morals, regulations related to property, law, state and society, agriculture and trade, history and stories, and religions, using the KNN, SVM, and Random Forest. This study uses data in English. By conducting tests on all classes, it is hoped that a thorough understanding of the strengths and weaknesses of each method will be obtained in the context of complex and thematic classification of religious texts.

2. RESEARCH METHOD

2.1. Research Stages

Figure 1 illustrates the phases of the research process used in this study [13]. The data used came from the Qur'anic translation and was separated into two categories: training data and test data. The data went through the preprocessing, vectorizer, and data separation stages. Next, the KNN, SVM or Random Forest methods were applied, where each method produced an optimal model, which was then tested on the test data to obtain the classification performance.

2.2. Dataset

The dataset used in this study is an English translation of the Quran. Each verse in the data was labeled a topic by which it consists of 15 topic categories. according to the topics grouped into classes, namely (1) Arkanul Islam (C1), Iman (C2), Al-Qur'an (C3), Science (C4), Amal (C5), Dakwah (C6), Jihad (C7), Human and Community relations (C8), Akhlak (C9), Rules Relating to Property (C10), Matters Relating to the law (C11), Country and Society (C12), Agriculture and Trade (C13), History and Story (C14), and Religions (C15) [14]. The data used were multiclass [15]. A multiclass dataset representation of such data is shown in Table 1. Each verse can fall into several classes, depending on its content [16]. This is the basic concept of multiclass classification, which is used to classify data into several relevant categories or labels [17, 18].

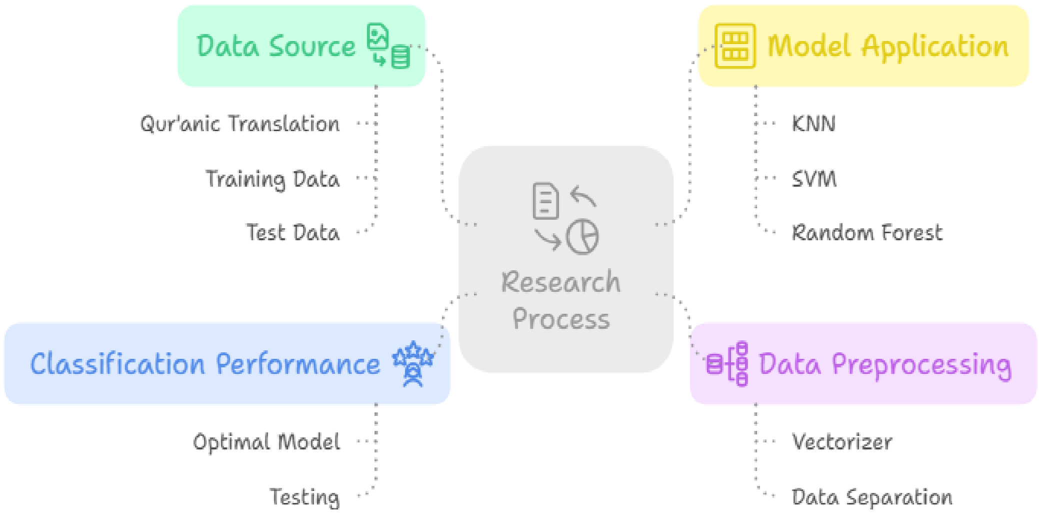


Figure 1. Research stages [13].

Table 1. Data representation multi-class.

Surah:Verse	Translation	Class (Label)
2:5	These are on a right course from their Lord and these it is that shall be successful.	Iman (C2), Amal (C5)
2:42	And do not mix up the truth with the falsehood, nor hide the truth while you know (it).	Matters relating to law (C11), religions(C15)
41:49	Man is never tired of praying for good, and if evil touch him, then he is despairing, hopeless.	Human and society (C8), Akhlak (C9)
4:68	And We would certainly have guided them in the right path.	Amal (C5), Religions (C15)
16:83	They recognize the favor of Allah, yet they deny it, and most of them are ungrateful.	Arkanul Islam (C1), human and community relations (C8)

2.3. Training and Validation Data

The dataset used in this study was divided into two main parts: training and testing data. Training data were used to conduct the research [13]. This division was carried out in an 80:20 ratio, where 80% of the data were used to train the model, and 20% of the data were used to test the model's performance after training. The dataset was randomly divided, into 4,988 verses used to train the model. Several variations of the model of the training process were tested on the validation data, which accounted for as much as 10% of the training data.

2.4. Data Testing

At this stage, tests are carried out on models that have been trained using data training to measure their performance on data that has never been seen before [12], namely, testing data as many as 1,248 verses. This testing process aims to assess how effective the model is in classifying the translated verses of the Qur'an into predetermined categories.

2.5. Text Preprocessing

Text preprocessing plays an important role in summarizing and grouping documents. This process includes several stages, such as tokenization, case folding, punctuation removal, stopwords, and stemming, all of which aim to prepare the data for easier processing by the model.

- Tokenization: The stage of breaking a text or sentence into small units such as words [19].
- Punctuation removal: This involves cleaning data from elements such as numbers, symbols, punctuation, links, hashtags, and usernames [15].
- Stopwords: Removing common words that do not significantly contribute to meaning, such as conjunctions or adverbs [19].
- Stemming: Simplify a word to its basic form by removing suffixes [20].

The entire preprocessing process was applied to both the training and test data. Effective preprocessing can significantly enhance the performance of classification models by reducing noise and improving the quality of input data [21].

2.6. Vectorizer

Feature construction or the formation of new features from raw data requires time and careful observation, as it is performed manually to ensure that the resulting features provide optimal benefits [22]. In this study, feature construction was performed using a vectorizer and was divided into two stages: feature extraction and feature selection. Feature extraction is where the system converts words into a vector that consists of two phases: primary extraction, which converts data into a standard format, and then converts it into a standard format. vector. Feature extraction can be performed using a representation bag of words, which is a commonly used NLP method [23]. Feature selection is a stage in data preprocessing that selects critical features in the analysis, improves model accuracy, and reduces computation time. This is performed using the TF. IDF and word count are techniques for measuring the value of word frequency in a document [24].

2.7. K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a method of classification of text or data that uses directed learning [9]. K-Nearest Neighbors (KNN) is a fundamental classification method and is often used in a variety of applications, including text classification [13]. KNN classifies each new data class from the nearest data k in the feature space. The algorithm is simple in concept and implementation, making it a good choice for baselineors when the computing resources are limited.

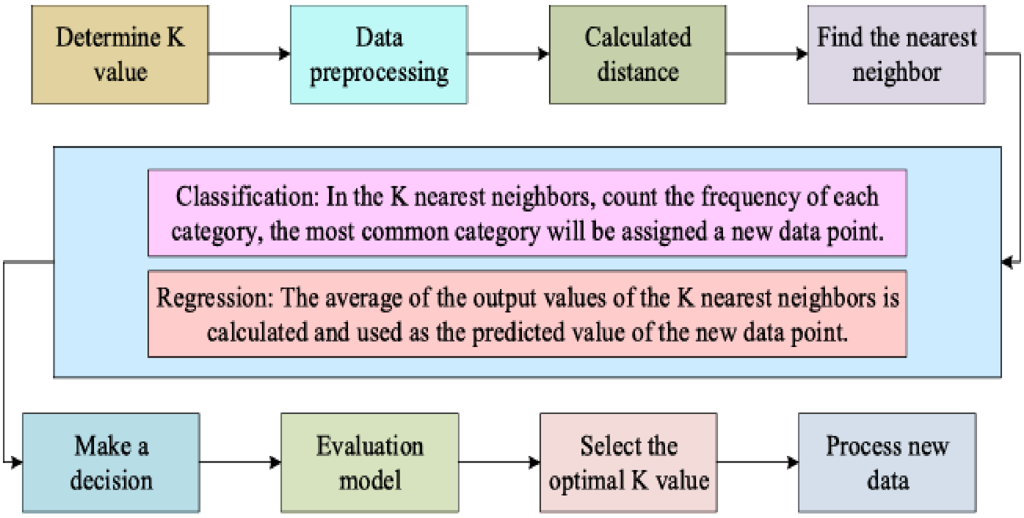


Figure 2. Stages of the KNN algorithm.

Figure 2 shows the stages of the KNN method. The KNN calculates the similarities between the unclassified, class-unknown samples and each sample in the training data, then sorts them and selects the top sample K. Subsequently, the class of the unknown sample was determined based on the class of the sample K. The category that appears most frequently among K samples is the category of the unknown sample [24].

2.8. Support Vector Machine (SVM)

Support Vector Machine (SVM) is a supervised learning method used to analyze data and recognize patterns in the data grouping process [25]. Support Vector Machine (SVM) is an algorithm used to group data with both linear and nonlinear patterns. An SVM can handle complex data because it uses the kernel concept, which is a way to move data to higher spaces to make it easier to separate. Using an SVM, different objects can be grouped according to their type. As a supervised learning method, SVM studies patterns from prelabeled data to predict classes from new data [5]. The SVM architecture is shown in Figure 2.

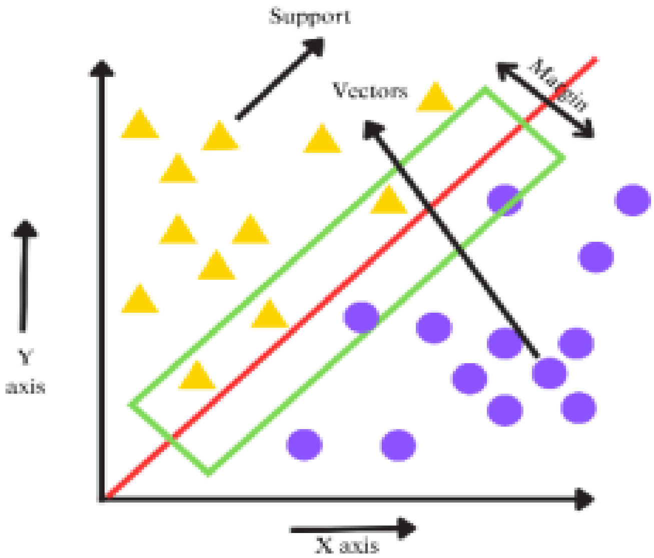


Figure 1 Algoritma support vector machine [26].

The SVM formula is expressed in Equation 1 as follows[27]:

$$f(x) = sign \sum_{i=1}^{n} (a_i y_i K(x_i x) + b) \quad (1)$$

information:
(x) : the data to be classified
x_i : Training data
y_i : Class labels
a_i : Weight
K(,x_ix) : a kernel that calculates the distance between and x_ix
b : bias

2.9. Random Forest

Random Forest is an ensemble-based algorithm developed using the Decision Tree method [28]. The ensemble algorithm itself is a combination of several machine-learning techniques to form a more powerful predictive model. Random Forest works by building a number of decision trees, and then the prediction results of each tree are collected, and most results (majority voting) are chosen as the final output [7], as shown in Figure 4.

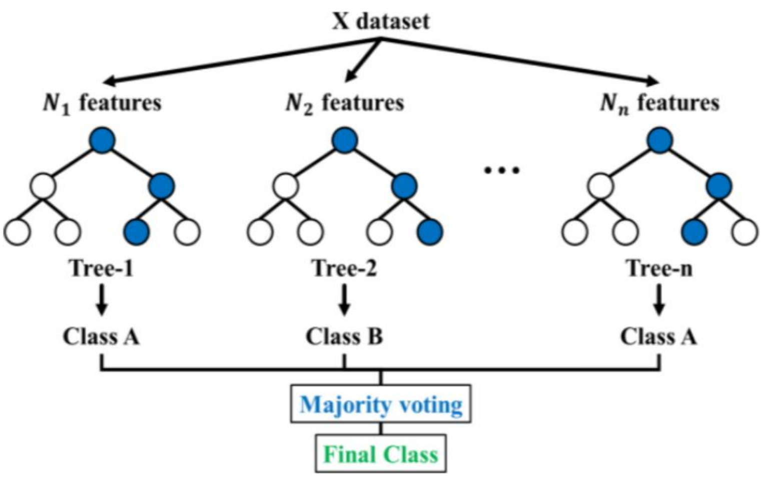


Figure 2. Random forest architecture [7].

This approach effectively addresses the drawbacks of using a single decision tree, which often results in inaccurate classification [29].

2.10. Evaluation

At this stage, tests were carried out on models that had been trained using data training to measure their performance on data that had never been seen before, namely, testing data as many as 1,248 verses. This testing process aims to assess how effective the model is in classifying the translated verses of the Qur'an into predetermined categories. The evaluation in this study used accuracy (1) and F1-Score (2).

$$Accuracy = \frac{\sum_{i=1}^{l} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}}{l} \times 100\% \tag{2}$$

$$F1 - Score = \frac{2 \times recall \times precision}{recall + precision} \tag{3}$$

information:

TP_i = True Positive is the amount of positive data correctly classified by the system for class i.

TN_i = True Negative which is the amount of negative data but is classified incorrectly by the system for the first class.

FP_i = False Positive, which is the amount of negative data that is classified incorrectly by the system for the first class.

FN_i = False Negative which is the amount of data that is positive but classified incorrectly by the system for the first class.

3. RESULTS AND DISCUSSION

In this study, a total of 6236 data were used. This study tested two text classification methods KNN, SVM, and RF. The evaluation was performed using an English Qur'an translation dataset with an 80:20 data-sharing scheme (80% training data and 20% test data), as well as full preprocessing for text data cleaning. The results of the KNN test are listed in Table 2.

Table 1. Result testing data training KNN.

Class	Train score		Val score	
	F1 (%)	Acc (%)	F1 (%)	Acc (%)
C1	57.26	92.20	53.30	92.99
C2	75.66	78.84	62.57	67.33
C3	58.63	92.54	49.86	91.18
C4	57.26	92.20	53.39	92.99
C5	58.59	89.69	54.41	90.18
C6	49.24	96.99	49.08	96.39
C7	53.49	94.63	52.23	94.79
C8	57.05	90.71	47.75	91.38
C9	56.15	89.13	50.49	86.57
C10	61.19	95.88	56.92	95.79
C11	53.28	96.97	49.49	98.00
C12	49.82	99.26	49.95	99.80
C13	49.80	99.20	49.86	99.44
C14	59.44	90.73	55.27	88.18
C15	52.79	95.37	49.13	96.59
Correspondence	56.64	92.96	52.25	92.11

Based on Table 2, the results of the data train the KNN method, where the method performs the preprocessing process. Next, test data are obtained, and the classification performance is produced, as shown in Table 3.

Table 2. Results of the KNN method test data.

Class	Data test	
	F1 (%)	Acc (%)
C1	57.52	93.03
C2	66.21	69.71
C3	54.32	91.51
C4	57.52	93.03
C5	54.45	88.94
C6	49.10	96.47
C7	53.59	94.31
C8	54.33	90.79
C9	50.53	87.10
C10	56.70	95.67
C11	48.96	95.91
C12	49.86	99.44
C13	49.89	99.55
C14	56.27	90.30
C15	51.75	95.03
Average	54.07	92.05

Table 3 shows the results of the testing data tests from all classes using the KNN method, which consists of F1-score values and accuracy data tests. Next, we tested the Support Vector Machine method as shown in Table 4.

Table 4. SVM method data train test results.

Class	Train score		Val score	
	F1 (%)	Acc (%)	F1 (%)	Acc (%)
C1	99.69	99.87	49.03	58.32
C2	99.80	99.86	45.69	47.70
C3	100.00	100.00	52.14	91.58
C4	100.00	100.00	52.14	91.58
C5	99.95	99.95	54.41	90.18
C6	100.00	100.00	54.39	96.59
C7	100.00	100.00	52.23	94.79
C8	99.97	99.97	47.64	90.98
C9	99.99	99.99	50.37	86.37
C10	99.93	99.95	52.82	95.39
C11	100.00	100.00	49.49	98.00
C12	99.66	99.94	49.90	99.60
C13	100.00	100.00	49.89	99.55
C14	99.92	99.92	49.40	86.97
C15	99.96	99.97	54.74	96.79
Correspondence	99.92	99.96	50.95	88.29

Based on Table 4, the results of the SVM method obtained a fairly good score. Next, test data are obtained, and the classification performance is produced, as shown in Table 5.

Table 5 shows the result of testing data tests from all classes using the SVM method, which consists of F1-score values and accuracy data tests. Next, testing was performed using the Random Forest method, as shown in Table 6.

It can be seen in Table 7 that the test data is carried out and produces classification performance.

Table 7 shows that the RF model is generally reliable in predicting the correct label. Table 8 shows a comparison of F1 scores from the three methods for metric confusion in Iman class (C2) test data.

Table 5. SVM method test data results.

Class	Data test	
	F1 (%)	Acc (%)
C1	48.56	61.46
C2	45.52	47.92
C3	54.94	91.43
C4	54.94	91.43
C5	50.26	88.62
C6	51.35	96.55
C7	48.45	93.99
C8	52.87	90.62
C9	48.21	86.78
C10	51.89	95.19
C11	48.94	95.83
C12	49.86	99.44
C13	49.90	99.60
C14	51.16	89.10
C15	54.54	95.11
Average	50.76	88.20

Table 6. RF method data train test results.

Class	Train score		Val score	
	F1 (%)	Acc (%)	F1 (%)	Acc (%)
C1	99.63	99.64	65.73	66.70
C2	99.76	99.78	65.91	70.49
C3	100.00	100.00	61.22	92.54
C4	99.83	99.94	60.40	92.54
C5	99.80	99.92	58.06	88.98
C6	100.00	100.00	55.93	97.10
C7	99.87	99.97	54.59	94.99
C8	99.85	99.94	57.76	90.53
C9	99.87	99.94	54.48	87.86
C10	99.84	99.97	70.30	96.10
C11	99.76	99.97	52.03	96.44
C12	98.88	99.97	49.72	98.89
C13	100.00	100.00	49.89	99.55
C14	99.85	99.94	63.06	90.65
C15	99.85	99.97	58.30	96.10
Correspondence	99.79	99.93	58.49	90.63

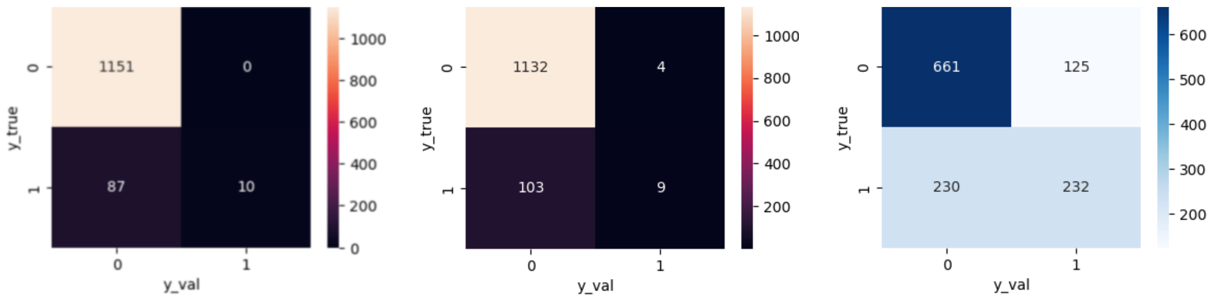
Table 7. Results of RF method test data.

Class	Data test	
	F1 (%)	Acc (%)
C1	67.36	68.91
C2	67.74	71.55
C3	60.88	91.83
C4	65.21	93.43
C5	52.14	88.78
C6	51.35	96.55
C7	53.59	94.31
C8	54.97	90.79
C9	55.99	87.50
C10	73.84	96.71
C11	48.94	95.83
C12	49.86	99.44
C13	49.90	99.60
C14	64.23	91.27
C15	61.19	95.59
Average	58.48	90.81

Table 8. F1 score data test for science and amal (C4 & C5) classes.

Class	KNN	SVM	RF
Science	57.52	54.94	65.21
Amal	54.45	50.26	52.14

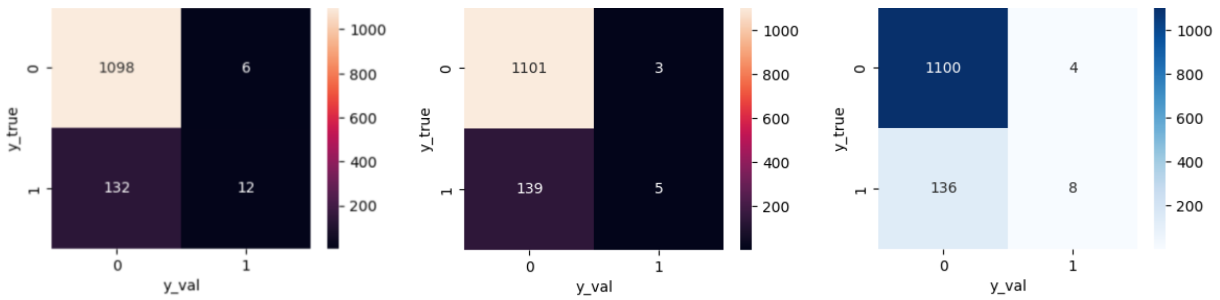
The Confusion Matrix provides a visual overview of the performance of the classification model. By comparing the matrix confusion of the optimal model, we can observe the difference in the number of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) [21].



(a) (b) (c)

Figure 3. Confusion metrics of the science class: (a) KNN; (b) SVM; and (c) RF.

Based on Figure 5, the confusion matrix shows the results of testing the three classification methods: KNN, SVM, and RF on Science class data. Each matrix displays the number of true and false predictions for each class.



(a) (b) (c)

Figure 6. Confusion metrics of the amal class: (a) KNN; (b) SVM; and (c) RF.

Figure 6 shows the confusion matrix of the three classification methods for the Amal class. The KNN method managed to correctly predict 12 Amal class data, while SVM 5 data, and RF 8 data. The amount of data that is not a Charity class and is successfully predicted correctly by the KNN is 1098, by SVM 1101, and by RF 1100. The difference between SVM and RF in predicting non-charity class data is only 1 data.

4. CONCLUSION

All three methods showed quite good results in classifying the topic of Qur'an translation. The Random Forest method provided the most balanced results, with an average F1-score of 58.48% and an accuracy of 90.81%, demonstrating its ability to recognize different classes well. KNN recorded the highest accuracy of 92.05%, but its F1 score was 54.07%, indicating that it tends to be more accurate in classes that appear frequently. Meanwhile, SVM had the lowest results with an F1-score of 50.76% and an accuracy of 88.20%, indicating that its performance was less stable than that of the other two methods, making it less effective than the other two methods in this classification task.

This research demonstrates that with proper preprocessing and vectorization, KNN, SVM, and Random Forest can all be effective for classifying translations of the Al-Quran, each with its own strengths and limitations. However, some suggestions can be considered for further research. First, it is recommended to explore other machine-learning methods, such as Neural Networks or Gradient Boosting, which might provide better performance in the context of text classification. Second, further research may consider the use of more advanced Natural Language Processing techniques, such as word embeddings or transformer models, to improve data representation and classification accuracy. In addition, it is important to perform a more in-depth analysis of the hyperparameter-tuning parameters for each model to determine the optimal configuration. Finally, subsequent research could expand the scope of the data by including translation variations from various sources to improve the generalization of the model. With these steps, it is hoped that the results of this research will be more useful and applicable in the context of understanding and analyzing the Quran text.

ACKNOWLEDGMENTS

We sincerely thank our colleagues and mentors at Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia, for their valuable insights and constructive feedback during this research on translation classification. We are also grateful to the anonymous reviewers and the editorial team for their helpful comments, which improved the final version of this article.

REFERENCES

[1] Gaanoun, K. & Alsuhaibani, M. (2025). Sentiment Preservation in Quran Translation with Artificial Intelligence Approach: Study in Reputable English Translation of the Quran. *Humanities and Social Sciences Communications*, **12**(1), 1–15.

[2] Bashir, M. H., Azmi, A. M., Nawaz, H., Zaghouani, W., Diab, M., Al-Fuqaha, A., & Qadir, J. (2023). Arabic Natural Language Processing for Qur'anic Research: A Systematic Review. *Artificial Intelligence Review*, **56**(7), 6801–6854.

[3] Al Salem, M. N., Altakhaineh, A. R., Alghazo, S., & Rayyan, M. (2024). On the Translation of Attributes of Allah in the Holy Quran: A Comparative Study. *Forum Linguist. Stud.*, **7**(1), 232–243.

[4] Choirulfikri, M. R., Lhaksamana, K. M., & Al Faraby, S. (2022). A Multi-Label Classification of Al-Quran Verses Using Ensemble Method and Naïve Bayes. *Building of Informatics, Technology and Science (BITS)*, **3**(4), 473–479.

[5] Zafar, A. & Iqbal, A. (2020). Machine Reading of Arabic Manuscripts Using KNN and SVM Classifiers. *Proceedings of the 7th International Conference on Computing for Sustainable Global Development, INDIACom 2020*, 83–87

[6] Hadi, W. M., Eljinini, M. A. H., & Alhawari, S. (2010). The Automated Arabic Text Categorization Using SVM and KNN. *Knowledge Management and Innovation: A Business Competitive Edge Perspective - Proceedings of the 15th International Business Information Management Association Conference, IBIMA 2010*, **2**, 757–763.

[7] Saputro, D. R. S. & Sidiq, K. (2023). Cable News Network (CNN) Articles Classification Using Random Forest Algorithm with Hyperparameter Optimization. *BAREKENG: Jurnal Ilmu Matematika Dan Terapan*, **17**(2), 847–854.

[8] Putra, O. V., Elfianita, F. D., Harmini, T., Trisnani, A., & Nugroho, A. (2022). iqurnet: A Deep Convolutional Neural Network for Text Classification on the Indonesian Holy Quran Translation. *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*, 86–91.

[9] Adeleke, A. O., Samsudin, N. A., Mustapha, A., & Nawi, N. M. (2017). Comparative Analysis of Text Classification Algorithms for Automated Labelling of Quranic Verses. *International Journal on Advanced Science, Engineering and Information Technology*, **7**(4), 1419–1427.

[10] Adeleke, A., Samsudin, N., Mustapha, A., & Khalid, S. A. (2019). Automating Quranic Verses Labeling Using Machine Learning Approach. *Indonesian Journal of Electrical Engineering and Computer Science*, **16**(2), 925–931.

[11] Rostam, N. A. P. & Malim, N. H. A. H. (2021). Text Categorisation in Quran and Hadith: Overcoming the Interrelation Challenges Using Machine Learning and Term Weighting. *Journal of King Saud University-Computer and Information Sciences*, **33**(6), 658–667.

[12] Ezz, M., Sharaf, M. A., & Hassan, A. A. A. (2019). Classification of Arabic Writing Styles in Ancient Arabic Manuscripts. *International Journal of Advanced Computer Science and Applications*, **10**(10).

[13] Alyasiri, O. M. & Cheah, Y. N. (2025). Multi–Class Text Classification using Machine Learning Techniques. *Engineering, Technology & Applied Science Research*, **15**(3), 22598–22604.

[14] Pane, R. A., Mubarok, M. S., & Huda, N. S. (2018). A Multi–Lable Classification on Topics of Quranic Verses in English Translation Using Multinomial Naive Bayes. *2018 6th International Conference on Information and Communication Technology (ICoICT 2018)*, 481–484.

[15] Mediamer, G., Adiwijaya, & Faraby, S. A. (2019). Development of Rule-Based Feature Extraction in Multi-label Text Classification. *Int. J. Adv. Sci. Eng. Inf. Technol.*, **9**(4), 1460–1465.

[16] Nurfikri, F. S. (2019). A Comparison of Neural Network and SVM on the Multi-Label Classification of Quran Verses Topic in English Translation. *Journal of Physics: Conference Series*, **1192**(1), 012030.

[17] Adeleke, A., Azah Samsudin, N., Hisyam Abdul Rahim, M., Kamal Ahmad Khalid, S., & Efendi, R. (2021). Multi-Label Classification Approach for Quranic Verses Labeling. *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, **24**(1), 484–490.

[18] Pal, K. & Patel, B. V. (2020). Automatic multiclass document classification of Hindi poems using machine learning techniques. *2020 International Conference for Emerging Technology (INCET)*, 1–5.

[19] Hamed, S. K. & Ab Aziz, M. J. (2018). Classification of Holy Quran Translation Using Neural Network Technique. *Journal of Engineering and Applied Sciences*, **13**(12), 4468–4475.

[20] Ahmed, M. A., Baharin, H., & Nohuddin, P. E. (2023). K-Means Variations Analysis for Translation of English Tafseer Al-Quran Text. *International Journal of Electrical and Computer Engineering (IJECE)*, **13**(3), 3255.

[21] Shrivash, B. K., Verma, D. K., & Pandey, P. (2025). A Novel Framework for Text Preprocessing Using NLP Approaches and Classification using Random Forest Grid Search Technique for Sentiment Analysis. *Economic Computation & Economic Cybernetics Studies & Research*, **59**(2).

[22] Azim, K., Tahir, A., Shahroz, M., Karamti, H., Vazquez, A. A., Vistorte, A. R., & Ashraf, I. (2025). Ensemble Stacked Model for Enhanced Identification of Sentiments from IMDB Reviews. *Scientific Reports*, **15**(1), 13405.

[23] Ishaq, M., Asim, M., Ahadullah, Cengiz, K., Konecki, M., & Ivković, N. (2025). Development of Novel Urdu Language Part of Speech Tagging System for Improved Urdu Text Classification. *Social Network Analysis and Mining*, **15**(1), 58.

[24] Xu, Q. (2025). Application of an Intelligent English Text Classification Model with Improved KNN Algorithm in the Context of Big Data in Libraries. *Systems and Soft Computing*, **7**, 200186.

[25] Ding, Z., Wang, Z., Zhang, Y., Cao, Y., Liu, Y., Shen, X., Tian, Y., & Dai, J. (2025). Trade-Offs Between Machine Learning and Deep Learning for Mental Illness Detection on Social Media. *Scientific Reports*, **15**(1), 14497.

[26] Maggioni, F. & Spinelli, A. (2025). A Novel Robust Optimization Model for Nonlinear Support Vector Machine. *European Journal of Operational Research*, **322**(1), 237–253.

[27] Asha, M. M., & Ramya, G. (2024). Artificial Flora Algorithm Based Feature Selection with Support Vector Machine for Cardiovascular Disease Classification. *IEEE Access*, **13**, 7293–7309.

[28] Sarawan, K., Polpinij, J., Somprasertsri, G., Rojarath, A., & Luaphol, B. (2025). Multiclass Classification Approach for Detecting Software Bug Severity Level from Bug Reports. *ICIC Express Letters. Part B, Applications: An International Journal of Research and Surveys*, **16**(5), 567–576.

[29] Afianto, M. F. & Al-Faraby, S. (2018). Text Categorization on Hadith Sahih Al-Bukhari Using Random Forest. *Journal of Physics: Conference Series*, **971**(1), 012037.