Institute of Research and Publication Indonesia (IRPI)
Public Research Journal of Engineering, Data Technology and Computer Science
Journal Homepage: https://journal.irpi.or.id/index.php/predatecs
Vol. 1 Iss. 2 January 2024, pp: 78-84
ISSN(P): 3024-921X | ISSN(E): 3024-8043

Text Classification of Translated Qur'anic Verses Using Supervised Learning Algorithm

Dhea Ananda[1*], Syahida Nurhidayarnis[2], Tiara Afrah Afifah[3], Muhammad Anang Ramadhan[4], Ilvan Mahendra[5]

[1,2,3]Department of Information Systems, Faculty of Science and Technology, Indonesia
[4]Puzzle Research Data Technology, Faculty of Science and Technology,
Universitas Islam Negeri Sultan Syarif Kasim Riau, Indonesia
[5]Study Program of Tafsir, Faculty of Ushuluddin, Al-Azhar University, Egypt

E-Mail: [1]12050322994@students.uin-suska.ac.id,
[2]12050321868@students.uin-suska.ac.id, [3]12050322968@students.uin-suska.ac.id,
[4]anang.ramadhan@gmail.com, [5]mahendrailvan8@gmail.com

Received Aug 7th 2023; Revised Dec 2nd 2023; Accepted Dec 26th 2023
Corresponding Author: Dea Ananda

Abstract

The Quran, comprising Allah's absolute divine messages, serves as guidance. Although reading the Quran with tafsir proves beneficial, it may not offer a comprehensive understanding of the entire message conveyed by the Al-Quran. This is due to the Quran addressing diverse topics within each surah, necessitating readers to reference interconnected verses throughout the entire chapter for a holistic interpretation. However, given the extensive and varied verses, obtaining accurate translations for each verse can be a complex and time-consuming endeavor. Therefore, it becomes imperative to categorize the translated text of Quranic verses into distinct classes based on their primary content, utilizing Fuzzy C-Means, Random Forest, and Support Vector Machine. The analysis, considering the obtained Davies-Bouldin Index (DBI) value, reveals that cluster 9 emerges as the optimal cluster for classifying QS An-Nisa data, exhibiting the lowest DBI value of 4.30. Notably, the Random Forest algorithm demonstrates higher accuracy compared to the SVM algorithm, achieving an accuracy rate of 66.37%, while the SVM algorithm attains an accuracy of 50.56%.

Keyword: Al-Quran, Classification, Fuzzy C-Means, Random Forest, Support Vector Machine (SVM)

1. INTRODUCTION

The Qur'an is the book of Allah revealed to Muhammad SAW to guide people from darkness to light [1], [2]. The Quran as a guide is also the guidance of the word of God, containing divine messages whose truth is absolute so that it becomes the main reference in living life and making decisions in accordance with His will.[3]. It can be likened to an ocean of unfathomable depth, and is the largest and most famous book in the world, consisting of 114 chapters and 30 sections. [4]. The Quran is also a source of law that regulates all aspects of life, both in worship and muamalah. By following the instructions contained therein, humans can achieve happiness and success both in this world and in the hereafter.

Qur'anic interpretation itself comprises different methods and approaches. Different interpretive approaches have been taken by scholars of Quranic sciences such as literary, juridical, philosophical, social, moral, historical approaches, and others. [5]. Interpreting means explaining, revealing, and interpreting, or revealing the meaning intended by the text, by its signs, or by its purpose [6]. Reading the Quran with tafsir is quite helpful but does not give a complete picture of what the book of the Quran wants to convey to its readers. This is because the Quran covers various topics in each surah, in order to interpret the Quran as a whole, the reader must refer to each verse in all interconnected surahs [7]. In today's digital age, many apps and platforms provide translations of Qur'anic verses in various languages. However, with such a large number and variety of verses, finding the right and accurate translation for each verse can be a time-consuming and complicated task.

Text classification is a subset of the topic of supervised learning that is applied to categorize data into different categories. [8]. The development of advances in technology today makes the classification of Al-Qur'an translation texts can be implemented utilizing an algorithm of machine learning. Machine learning is an AI implementation that applies mathematical approaches to build automatic models from a set of data, with the objective of allowing the computer to "learn" [9]. In the context of Qur'anic translation text classification, supervised learning algorithms can learn specific linguistic patterns and language structures in

Qur'anic verses. By utilizing features extracted from the training data, such as key words, sentence structure, or statistical methods, the model can classify new unknown texts into appropriate translation categories. This can be done using Fuzzy C-Means, Random Forest and Support Vector Machine (SVM) algorithms.

Fuzzy C-Means, Random Forest, and Support Vector Machine (SVM) algorithms are effective supervised learning algorithms in classification. The use of these algorithms, such as SVM which achieves high accuracy, FCM as a superior fitness function, and Random Forest which improves prediction accuracy, has been proven in previous studies [10], [11], [12]. Research conducted by M. Al-Nokrashy et al. proved the superiority of FCM in GA-based classification with an average accuracy of 87.84% [13], while improvements to the Random Forest model by Thulasi Bikku and K. P. N. V. Satyasree improved prediction accuracy and software quality [14]. Recent research by Mohamed and Behaidy, using logistic regression, random forest, and SVM, showed that ensemble modeling achieved average hamming loss, recall, precision, and F1-Score of 0.224, 81%, 75%, and 77% [15].

Based on this background, the problem found is that there is no classification of the translated text of the Qur'an verse according to the main content of the book. So it is necessary to classify the translated text of the Qur'an verse into several classes according to the main content of the verse. This research applies Fuzzy C-Means, Random Forest and Support Vector Machine models to measure the best performance of the model. Text classification aims to produce text classes of Al-Qur'an translation verse documents according to the content of the verse in the Al-Qur'an.

2. MATERIAL AND METHOD

Before collecting data, the most basic stage that is important to do is literature study, this is done to help the preparation process by collecting information from various books and journals related to the problems discussed. Next, preprocessing the data that has been collected and carried out and continued with weighting, grouping data using Fuzzy C-Means, and classifying by looking at the accuracy of the algorithms used, namely Random Forest and Support Vector Machine (SVM). The last stage carried out is making conclusions from the results of the research that has been done.

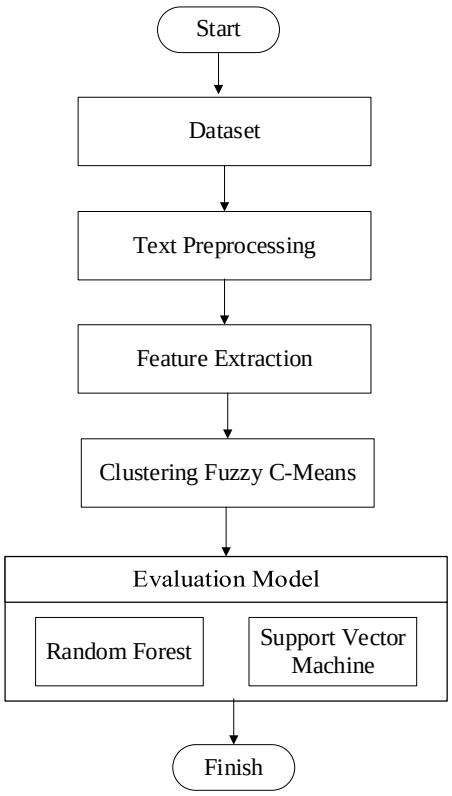

Figure 1. Research Methodology

2.1 Fuzzy C-Means

Fuzzy clustering is a fairly novel trend in clustering data, where a sample does not exclusively belong to one cluster but becomes a member of any cluster with a degree of membership. Because of its versatility and resilience to uncertainty, the most notable fuzzy clustering algorithm is Fuzzy C-Means (FCM). This algorithm was unveiled by Jim Bezdek in 1981 [16]. The equations contained in Fuzzy C-Means are:

1. Calculate the cluster center with the formula 1.

$$V_{kj} = \frac{\sum_i^n((U_{ik})^w X_{ij})}{\sum_{i=1}^n (U_{ik})^w} \quad (1)$$

2. Calculate the objective function at the tth iteration with the formula 2.

$$P_t = \sum_{i=1}^n \sum_{k=1}^c ([\sum_{j=i}^m (X_{ij} - V_{kj})](U_{ik})^w \quad (2)$$

3. Calculate the change in partition matrix Uik using the formula3.

$$U_{ik} = \frac{\sum_{j=1}^m (X_{ij} - V_{kj})^2]^{\frac{-1}{w-1}}}{\sum_{k=1}^c [\sum_{j=1}^m (X_{ij} - V_{kj})^2]^{\frac{-1}{w-1}}} \quad (3)$$

2.2 Random Forest

Random Forest was originally invented by fellow American computer scientist Tin Kam Ho in 1995 and developed by California Academy of Sciences member Breiman. It's a classification that contains a tree classification structure where each tree generates voting units for popular class inputs. Simply put, Random Forest comprises a collection of several decision trees, which are used to classify the data. [17]. Random Forest required zero parameter tuning. A tree-like method like random forest is built from samples from the data set, selecting fewer features, and finding the values that make the best separation in our data set [18].

Furthermore, a simple algorithm description of Random Forest can be seen as follows :

1. Formation of Decision Tree as equation 4.

$$E(S) = \sum_{i=1}^c - \ p_i \log_2 p_i \quad (4)$$

2. Formation of Random Forest
   The split that the tree uses to partition a node into its two descendants is chosen by considering every possible split on each predictor variable and selecting the "best" one according to some criteria. In regression, if the response values at a node are y1, ..., yn, a typical splitting criterion is the mean square at the node expressed as equation 5.

$$Q = \frac{1}{n} \sum_{i=1}^n \ (y_i - \bar{y})^2 \quad (5)$$

   where:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n 1 \ y_i$$

2.3 Support Vector Machine (SVM)

SVM is supervised model learning with a corresponding learning architecture that recognizes data utilized for both classification and regression analysis [19]. SVM is employed in cases where there are precisely two classes of data. SVM classify data by discovering the optimal hyperplane that decouples all data points from each class [20]. The distinguishing feature compared to other classifiers is that SVM search for a solution that reduces the probability of incorrect classification to a minimum. During the past decade, SVM have been applicable to a wide range of fields: Recognizing various characteristics of individuals including pictures, objects, sounds, fingerprints, and handwriting [21].

A commonly used model for solving classification and regression problems is the linear model, which is a linear combination of basis functions [21].

$$y(x,w) = \sum_0^{M-1} w_j \emptyset_j(x) = \ w^T \emptyset(x) \quad (6)$$

Where: $x = (x_1, x_2, ..., x_D)^T$ is variabel input, $w = (w_0, w_1, ..., w_D)^T$ is parameter, $\emptyset(x)$ is basis function, and M is total number of parameters of the model.

3. RESULTS AND DISCUSSION

3.1 Data Coollection

The initial stage in this research is collecting data. The type of data used in this research is text. The data used as the object of research is Q.S An-Nisa: 4 which is obtained from the Ministry of Religion's Al-Quran. The data used in the form of translation of Q.S An-Nisa:4 in Indonesian language taken from verse 1 to verse 176 and shown in Table 1.

Table 1. Initial Data

Verse	Translation
1	Wahai manusia! Bertakwalah kepada Tuhanmu yang...
2	Dan berikanlah kepada anak-anak yatim (yang su...
3	Dan berikanlah maskawin (mahar) kepada perempu...
4	Dan janganlah kamu serahkan kepada orang yang...
...	...
175	Adapun orang-orang yang beriman kepada Allah d...
176	Mereka meminta fatwa kepadamu (tentang kalalah...

3.2 Preprocessing Data

The next stage is preprocessing which consists of cleaning and tokenzing, stopwords, stenming and converting words into sentences. The cleaning process is usually used to clean data that is considered meaningless or has no value such as hashtags, emojis, etc. In the tested data, the cleaning process is listed in Table 2 where the Quran text is cleaned from unnecessary characters or symbols, such as punctuation marks, double dashes, or special notations.

Table 2. Cleaning of punctuation

Verse	Translation
166	sesungguhnya orang orang yang kafir dan mengha...
167	sesungguhnya orang orang yang kafir dan melaku...
168	kecuali jalan ke neraka jahanam mereka kekal...
168	wahai manusia sungguh telah datang rasul mu...
...	...
175	adapun orang orang yang beriman kepada allah d...
176	mereka meminta fatwa kepadamu tentang kalalah...

After cleaning, the next step is tokenizing which is an important step in text processing that allows us to organize, analyze, and manipulate text more effectively as shown in Table 3.

Table 3. Tokenizing

Verse	Translation
166	[sesungguhnya, orang, orang, yang, kafir, dan,...
167	[sesungguhnya, orang, orang, yang, kafir, dan,...
168	[kecuali, jalan, ke, neraka, jahanam, mereka,...
169	[wahai, manusia, sungguh, telah, datang, rasul,...
...	...
175	[adapun, orang, orang, yang, beriman, kepada, allah,...
176	[mereka, meminta, fatwa, kepadamu, tentang,...

The last stage of data preprocessing in this research consists of 2 processes, namely stopwords and stemming. Stopwords are common words that always appear in text but tend not to provide much information or special meaning in text processing. The stopword results can be seen in Table 4.

Table 4. Stopword

Verse	Translation
1	[manusia, bertakwalah, tuhanmu, menciptakan, a...
2	[berikanlah, anak, anak, yatim, dewasa, harta,...
3	[berikanlah, maskawin, mahar, perempuan...
4	[serahkan, orang, sempurna, akalnya, harta, ke...
5	[ujilah, anak, anak, yatim, umur, menikah, pen...
6	[laki, laki, hak, harta, peninggalan, orang, t...
...	...

Stemming is one of the techniques used in the pre-processing stage of text in data mining to reduce words to their basic form or root word. The goal is to eliminate variations in word forms that may occur in the text, so that words that have the same root word can be treated as the same entity. The stemming results can be seen in Table 5.

Table 5. Stemming

Verse	Translation
1	[manusia, takwa, tuhan, cipta, adam...
2	[beri, anak, anak, yatim, dewasa, harta,...
3	[khawatir, laku, adil, hak, hak, perempuan,...
4	[beri, maskawin, mahar, perempuan, nikah,...
5	[serah, orang, sempurna, akal, harta, kuasa,...
...	...

3.3 Feature Extraction

An easy way to create layouts is to use this guide directly. Download from link provided. Based on the stemming results, then perform feature extraction. Feature extraction means calculating the weighting of words using the TF-IDF technique. Weighting is done based on the frequency of occurrence of words in each document. The TF-IDF performed has broken down all words into a column with the results of 631 words obtained in 176 rows. The results of the calculation of word weights using TF-IDF are shown in Table 6.

Table 6. TF-IDF Calculation Results

Verse	acuh	ada	adakan	Adam	adil	agama	...	zalim	zina
1	0.0	0.0	0.0	0.215464	0.000000	0.000000	...	0.000000	0.0
2	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.000000	0.0
3	0.0	0.0	0.0	0.000000	0.328177	0.000000	0.0	0.134526	0.0
4	0.0	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.000000	0.0
...	...	...	...	...	...	...	...	...	...
175	0.0	0.0	0.0	0.000000	0.000000	0.277146	...	0.000000	0.0
176	0.0	0.0	0.0	0.000000	0.000000	0.000000	...	0.000000	0.0

3.4 Fuzzy C-Means

The Fuzzy C-Means algorithm here is carried out to classify data, so that classification can be carried out afterwards using the Random Forest and Support Vector Machine (SVM) algorithms. In grouping the data here, testing is carried out in selecting the number of clusters using the Davies-Bouldin Index (DBI) where if the number of clusters obtained the lowest DBI value shows the best results. In Table 7 we can see that the best cluster results are obtained at 4.307558 with k = 9 trials.

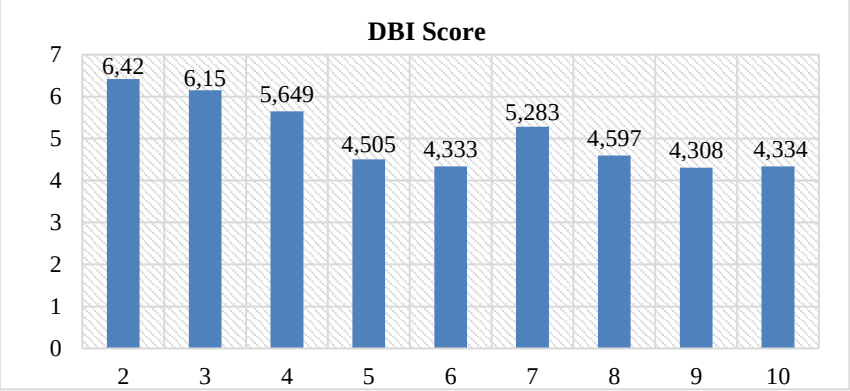

Figure 2. DBI Result

Based on the clustering results, further wordcloud visualization is done to find the most frequently occurring words in Q.S An-Nisa. The following visualization of wordcloud in cluster 9 is shown in Figure 3.

Figure 3. Wordcloud Q.S An-Nisa

3.5 Random Forest and Support Vector Machine (SVM) Classification

The clustering results that have been done before, found 9 clusters. So that 9 groups are obtained to be used as a labeling reference in conducting accuracy tests using random forest and SVM classification algorithms which can be seen in Table 8.

Table 8 Cluster labeling for classification

Verse	Translation	Cluster	Label
1	manusia takwa tuhan cipta adam allah cipta pas...	cluster9	9
2	beri anak anak yati dewasa harta tukar buruk...	cluster6	6
3	khawatir laku adil hak hak perempuan yatim bil...	cluster9	9
4	beri maskawin mahar perempuan nikah beri penuh...	cluster9	9
5	serah orang sempurna akal harta kuasa jadi all...	cluster4	4
...	...	...	...

In this research, classification uses two different algorithms, namely Random Forest and Support Vector Machine. These two algorithms were chosen based on certain considerations that refer to previous research on similar topics. The accuracy of an algorithm is influenced by several things. In this research, the data sharing technique used is accuracy. The results of the accuracy test conducted are shown in Figure 4.

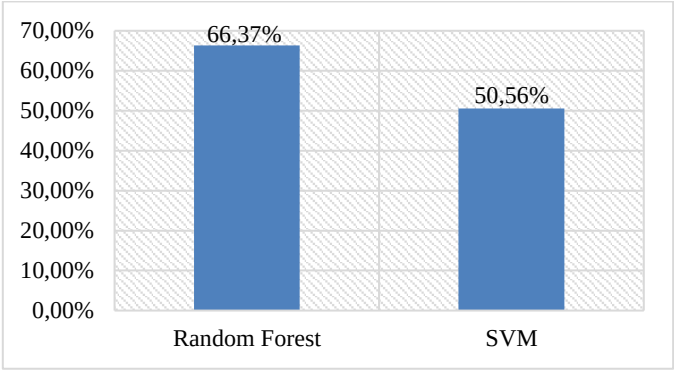

Figure 4. Accuracy Test Results

4. CONCLUSION

This research has classified the translated text of the Qur'an verses according to the main content of the book using random forest and svm algorithms. Based on the dbi value obtained, the result shows that cluster 9 is the best cluster for clustering QS An-Nisa data with a dbi value of 4.30. The random forest and svm algorithms provide good accuracy. However, the random forest algorithm has a higher accuracy than the svm algorithm. Where the accuracy of the random forest algorithm is 66.37% while the svm algorithm gets an accuracy of 50.56%. Thus, it can be concluded that text classification models using Random Forest and SVM algorithms have the potential to be used effectively but still have a lot of room for improvement. This research is still in the initial experimental stage. We still need additional experiments using other classification algorithms. For future research, parameter hypertuning can be applied to each algorithm to significantly improve accuracy.

REFERENCES

[1] N. S. Huda, M. S. Mubarok, and Adiwijaya, "A multi-label classification on topics of quranic verses (english translation) using backpropagation neural network with stochastic gradient descent and adam optimizer," *2019 7th Int. Conf. Inf. Commun. Technol. ICoICT 2019*, pp. 1–5, 2019, doi: 10.1109/ICoICT.2019.8835362.

[2] R. Umar and I. Ulumuddin, "Using of Exact Queries and Expansion Queries in Searching for Indonesian Translated Al-Quran Verses," *J. Mantik*, vol. 3, no. 2, pp. 10–19, 2019, [Online]. Available: http://iocscience.org/ejournal/index.php/mantik/article/view/882/595

[3] A. Aboamro and H. Rizapoor, "Unveiling the Divine Text : Exploring the Analytical Interpretation of the Holy Quran," no. 3, pp. 39–48, 2023.

[4] M. Mohammed *et al.*, "Machine Translated by Google Surat Ilmu Informasi Model Pembelajaran Mesin untuk Identifikasi Al-Qur ' an Reciter Memanfaatkan K-Nearest Neighbor dan Artificial Neural Jaringan Machine Translated by Google Surat Ilmu Informasi Model Pembelajaran Mesin unt," vol. 11, 2022.

[5] M. Sharifi, "An examination of the nature and necessity of feminist interpretation of the Holy Quran," *Kom Cas. za Relig. Nauk.*, vol. 9, no. 2, pp. 65–85, 2020, doi: 10.5937/kom2002065s.

[6] A. Songgirin, "Tafsīr Al-Quran Dengan Al-Quran," *Al Burhan J. Kaji. Ilmu dan Pengemb. Budaya Al-Qur'an*, vol. 21, no. 01, pp. 88–110, 2021, doi: 10.53828/alburhan.v21i01.221.

[7] M. H. Albar, F. A. Bachtiar, and Indriati, "Pengelompokan Terjemah Al-Quran Departemen Agama Menggunakan Metode Fuzzy C-Means," vol. 1, no. 1, pp. 1–10, 2020.

[8] F. S. Nurfikri and Adiwijaya, "A comparison of Neural Network and SVM on the multi-label classification of Quran verses topic in English translation," *J. Phys. Conf. Ser.*, vol. 1192, no. 1, 2019, doi: 10.1088/1742-6596/1192/1/012030.

[9] D. I. A. Putra and M. Yusuf, "Proposing machine learning of Tafsir al-Quran: In search of objectivity with semantic analysis and Natural Language Processing," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 1098, no. 2, p. 022101, 2021, doi: 10.1088/1757-899x/1098/2/022101.

[10] J. Laosai and K. Chamnongthai, "Acute leukemia classification by using SVM and K-Means clustering," *2014 Int. Electr. Eng. Congr. iEECON 2014*, pp. 1–4, 2014, doi: 10.1109/iEECON.2014.6925840.

[11] M. N. A. Al-hamadani, "Classification and analysis of the MNIST dataset using PCA and SVM algorithms," vol.71, no. 2, March, 2023, pp: 221-238, doi: 10.5937/vojtehg71-42689.

[12] F. S. Utomo, N. Suryana, and M. S. Azmi, "Stemming impact analysis on Indonesian Quran translation and their exegesis classification for ontology instances," *IIUM Eng. J.*, vol. 21, no. 1, pp. 33–50, 2020, doi: 10.31436/iiumej.v21i1.1170.

[13] M. Salah, "K-means versus fuzzy c-means as objective functions for Genetic Algorithms- based classification from aerial images and LIDAR data," no. July, 2017.

[14] T. Bikku, "A Boosted Random Forest Algorithm for Automated Bug Classification A Boosted Random Forest Algorithm," no. June, 2023, doi: 10.1007/978-981-99-0838-7.

[15] E. H. Mohamed and W. H. El-Behaidy, "An Ensemble Multi-label Themes-Based Classification for Holy Qur'an Verses Using Word2Vec Embedding," *Arab. J. Sci. Eng.*, vol. 46, no. 4, pp. 3519–3529, 2021, doi: 10.1007/s13369-020-05184-0.

[16] S. Zhou, D. Li, Z. Zhang, and R. Ping, "A New Membership Scaling Fuzzy C-Means Clustering Algorithm," *IEEE Trans. Fuzzy Syst.*, vol. 29, no. 9, pp. 2810–2818, 2021, doi: 10.1109/TFUZZ.2020.3003441.

[17] M. P. Utami, O. D. Nurhayati, and B. Warsito, "Hoax Information Detection System Using Apriori Algorithm and Random Forest Algorithm in Twitter," *6th Int. Conf. Interact. Digit. Media, ICIDM 2020*, no. Icidm, 2020, doi: 10.1109/ICIDM51048.2020.9339648.

[18] D. P. Mohandoss, Y. Shi, and K. Suo, "Outlier Prediction Using Random Forest Classifier," *2021 IEEE 11th Annu. Comput. Commun. Work. Conf. CCWC 2021*, pp. 27–33, 2021, doi: 10.1109/CCWC51732.2021.9376077.

[19] D. Valero-Carreras, J. Alcaraz, and M. Landete, "Comparing two SVM models through different metrics based on the confusion matrix," *Comput. Oper. Res.*, vol. 152, no. April 2022, p. 106131, 2023, doi: 10.1016/j.cor.2022.106131.

[20] D. Keerthana, V. Venugopal, M. K. Nath, and M. Mishra, "Hybrid convolutional neural networks with SVM classifier for classification of skin cancer," *Biomed. Eng. Adv.*, vol. 5, no. July 2022, p. 100069, 2023, doi: 10.1016/j.bea.2022.100069.

[21] E. Adıgüzel, N. Subaşı, T. V. Mumcu, and A. Ersoy, "The effect of the marble dust to the efficiency of photovoltaic panels efficiency by SVM," *Energy Reports*, vol. 9, pp. 66–76, 2023, doi: 10.1016/j.egyr.2022.10.358.