Research Article

Comparison of K-Nearest Neighbor and Support Vector Machine Methods in Sentiment Analysis of Offline Courses

[1]Muhammad Arrazi Alghifari Bimbe, [2]Jefri Marzal, [3,*]Ulfa Khaira

[1,2,3] Department of Information Systems, Jambi University, Indonesia

* Corresponding Author: ulfa.ilkom@gmail.com

Abstract: The implementation of the New Normal policy triggers various responses from the community regarding offline learning. Some students accept this system well and have a positive view towards the adaptation of new habits. To understand public opinion on this policy, sentiment analysis is used to categorize opinions into positive or negative categories. This method is very useful in extracting opinions from social media and analyzing public responses to an issue. In sentiment analysis, two methods that are often used are K-nearest neighbor (KNN) and Support Vector Machine (SVM). KNN classifies data based on the closest distance, but this method is quite susceptible to noise. In contrast, SVM works by determining the optimal hyperplane to separate data classes, making it more stable in classification. In this research, tests were conducted using a 90:10 split data scenario. The analysis shows that the accuracy of the Support Vector Machine is higher than K-Nearest Neighbors.SVM recorded an accuracy of 63.39%, while KNN only reached 38.80%. In addition, based on performance evaluation, SVM excels in Precision, Recall, and F1-Score aspects when compared to KNN. Based on these results, it can be concluded that in sentiment analysis related to offline learning after the New Normal policy, the Support Vector Machine method is more effective than the K-Nearest Neighbor, both in terms of accuracy and overall model performance.

Keywords: New Normal, Sentiment Analysis, Machine Learning, Online Learning, KNN

Citation: Bimbe, M. A. A., Marzal, J., & Khaira, U. (2025). Comparison of K-Nearest Neighbor and Support Vector Machine Methods in Sentiment Analysis of Offline Courses. *Iota*, 5(1). https://doi.org/10.31763/iota.v5i1.898

Academic Editor: Adi, P.D.P

Received: January 04, 2025

Accepted: February 15, 2025

Published: March 04, 2025

1. Introduction

The magnitude of the influence and benefits of Sentiment Analysis, causing research or applications regarding Sentiment Analysis to grow rapidly, even in America approximately 20-30 companies are using Sentiment Analysis to obtain information about public sentiment towards company services. Sentiment Analysis is a classification, but in reality, it is not as easy as the usual classification process because it is related to the use of language. There is ambiguity in the use of words, the absence of intonation in a text, and the development of the language itself.

To find out the best method that can be used in sentiment analysis, a comparative analysis of two methods is carried out, this is done because currently there are many studies on sentiment analysis. In doing sentiment analysis there are various methods such as Naive Bayes, Maximum Entropy, Deep Learning LSTM, K-Nearest Neighbor (KNN), and Support Vector Machine (SVM) [1,2,3,4].

This research will compare the K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) methods [5,6,7]. K-Nearest Neighbor is an algorithm that is easy to implement with a high level of effectiveness and is suitable for various problems related to classification. Meanwhile, Support Vector Machine is one of the algorithms that is suitable for text classification. The ability of the Support Vector Machine to find the best hyperplane gives this algorithm a high level of generality and makes it the algorithm with the best accuracy compared to other algorithms.

The sample data used is data on public comments on offline lectures after the new normal policy. The New Normal is a new life where people continue to carry out various activities as usual but still apply the health protocols set by the government so that the spread of the COVID-19 virus can be overcome. Along with the New Normal policy, the community (students) slowly began to be directed to take part in offline learning or lectures again. Offline learning can be defined as learning that is carried out outside of face-to-face by teachers and students but is carried out offline, which means that the teacher provides material in the form of hardcopy assignments to students and then carried out outside of school or lectures. This New Normal policy raises several opinions in the community, some agree with the implementation of offline learning and some disagree. Students' perceptions of offline classroom learning with the adaptation of new habits are positive or can be said to strongly agree. Because it is easier to understand the material in class, can do questions and answers.

In addition, students also enjoy the process of working on assignments given in class directly with the teacher. To find out the public response to the implementation of offline learning or public opinion about offline learning, there are social media that can be used. One of them that is popular at the moment is Twitter.

2. Literature Review

2.1 Text Mining

The purpose of text mining is to obtain useful information from a set of documents. So, the data source used in text mining is a set of text that has an unstructured or at least semi-structured format. The specific tasks of text mining include text categorization and text clustering.

2.2 Sentiment Analysis

The basic task of sentiment analysis is to focus on classifying the polarity of the text in the data whether the opinions expressed in the data are positive, neutral, or negative. Many methods can be used in performing sentiment analysis including naïve Bayes, Support Vector Machines, lexicon-based, K-nearest neighbor (K-NN), and so on.

2.3 Classification

Classification is a method of grouping or categorizing an unknown data object into a certain class based on existing classes. Classification in sentiment analysis usually solves problems by classifying them into 3 classes, such as positive, neutral, and negative.

2.4 Machine Learning

Machine learning aims to learn or recognize the structure of data through several patterns and convert the data into a model so that it can be used by many people [8,9,10]. The K-NN algorithm is a lazy learning method where no model is learned from testing data, so it only learns from test examples to be classified [11,12,13]. KNN calculates the closeness between new cases (test data) and old cases (training data) based on matching weights from several existing features. After the distance is calculated, the closest distance to the training data is considered to have similarities [14,15,16]. Support Vector Machine (SVM) is a supervised learning method that can analyze data and recognize classification patterns [17,18]. Support Vector Machine (SVM) is a supervised learning method that can analyze data and recognize classification patterns [19,20]. Figure 1 is an Illustration of the SVM method.

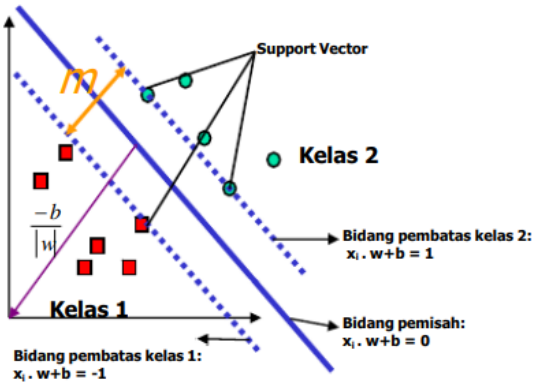

Figure 1. Illustration of the SVM method [15]

3. Method

This research will adopt methods and approaches that have been studied in previous case studies. The following is a framework that has been organized in such a way based on the methods used in Figure 2.

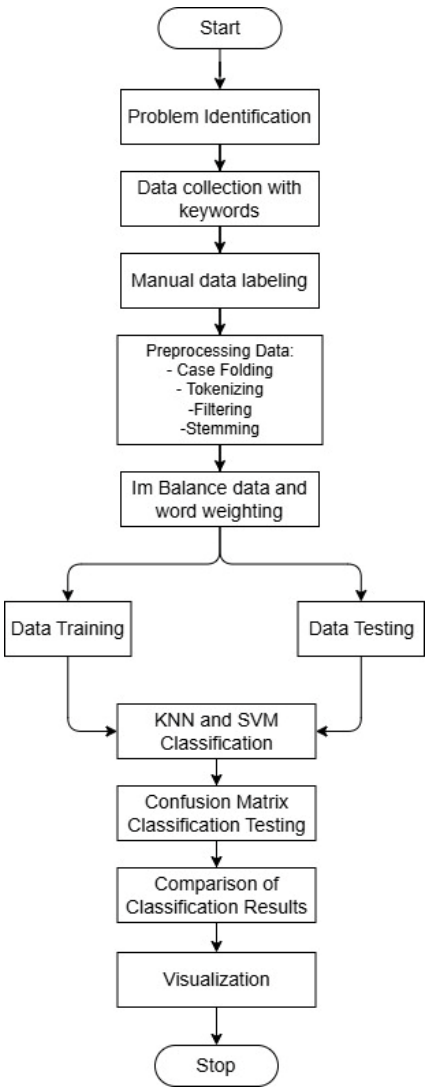

Figure 2. Flowchart System

4. Result and discussion

4.1 Data Scrapping & Labelling

The first stage in performing the sentiment analysis process is data collection. Data is taken from Twitter with a search query about "*Offline Lectures*" and then the URL from the search is inputted by exportcomment.com to be exported to Excel. Figure 3 is a Data retrieval from Twitter.

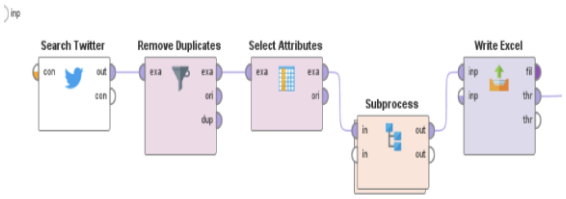

Figure 3. Data retrieval from Twitter

Moreover, the process of labeling the content is carried out to group the data into several categories that will be used by the author. The classes applied are data with positive and negative values. The purpose of this labeling process is to divide the data into two parts, namely testing data and training data. Testing data is data that will be used to test the results of training data, while training data is data used to analyze information in the system so that certain patterns can be recognized [14]. Here is an example of a dataset that has been labeled.

Table 1. Dataset label example

Text	Sentiment
It's great to be able to study offline again after a long time! Interaction with lecturers and friends is more effective. Hopefully, it will continue to be healthy and smooth!	Positive
It feels uncomfortable to study offline during conditions that are still vulnerable. Health protocols are often ignored, so I'm constantly worried.	Negative

After the data is labeled, the next step is the preprocessing stage, which is the process of preparing the data to be ready for analysis. This stage includes several processes, such as cleansing, duplicate removal, case folding, tokenization, stopword filtering, and stemming for Indonesians. Figure 4 is an Operator Pre-Processing.

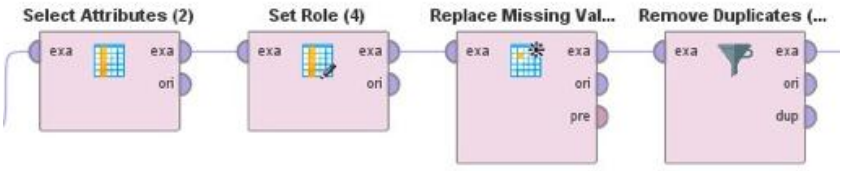

Figure 4. Operator Pre-Processing

With Google Colab processing tools using the Python programming language, here is an example of the results of the data cleaning that has been carried out:

Table 2. Data Cleaning Results

Text	Cleaned Text
Especially since college, I have never wanted to invite friends to pray again because my friends said "In the past, when we were in hybrid college, we prayed because it didn't feel good with you" And yes, since college offline no one has prayed at all 🤪👍	Especially since college, I have never wanted to invite friends to pray again because my friends said that in the past when we were in hybrid college we prayed because it didn't feel good with us and yes, since college offline no one has prayed at all.

4.2 Data Pre-Processing

The data preprocessing stage is carried out to convert data information from data sources that have been scraped previously into a standard format or more structured data because basically, the scraped data is still unstructured as previously explained. Some stages in data preprocessing include Case Folding, tokenizing, filtering, and stemming. After data preprocessing, clean data is obtained which will facilitate the classification process later. Figure 5 shows an example of Text Processing.

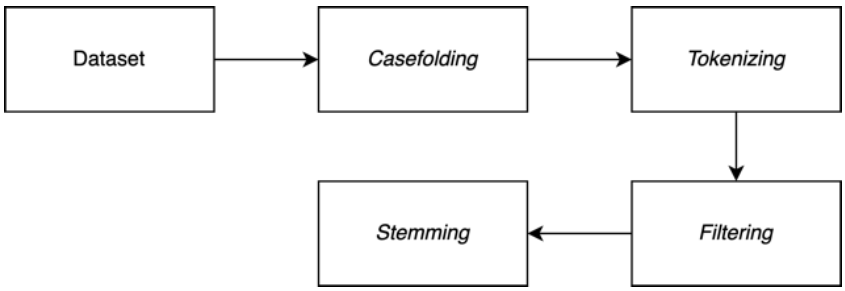

Figure 5. Text Processing

4.3 Case Folding

In writing tweets, there are usually different letterforms, this stage is the process of letter uniformity, related to capital letters or not. The following is an example of a case folding sentence, the input "PSBB effect lockdown flavor directly lost trillions", becomes the output "*PSBB* effect lockdown flavor directly lost trillions".

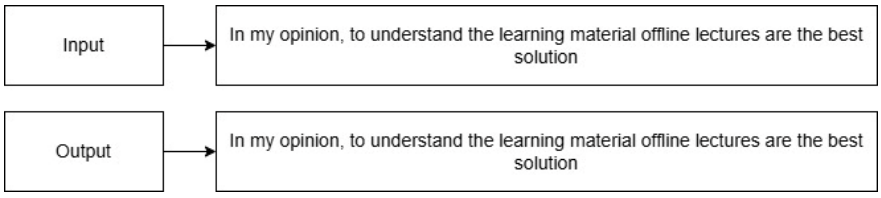

Figure 6. Case Folding Process

4.4 Filtering

Filtering is the stage of removing words that appear in large numbers but are considered meaningless (stopwords). A list of stopwords is a set of words that are widely used in various languages. The reason for removing words related to text mining is that they are too common so that users can focus on other words that are much more important.

Table 3. Example of Stopword

Rasa	*Lockdown*	*Kalau*	*Pada*	*Yaitu*
Aku	*Dia*	*Kami*	*Saja*	*Asik*
Bapak	*Ini*	*Lalu*	*Hai*	*Hari*
Berbagai	*Itu*	*Lewat*	*Untuk*	*Masa*
Cara	*Jadi*	*Meski*	*Yang*	*Tapi*

4.5 Tokenizing

Tokenization is the process of sorting and separating a sentence into several words, called tokens. In certain sentences, this process can also eliminate unnecessary punctuation so that it will facilitate the data processing process on Rapidminer [14]. There are tokenization models that can be used including, unigram, bigram, trigram, and ngram. The following are the results of the tokenizing process that has been carried out.

Table 4. Tokenizing Process

Tokenizing	**Stemming**
apalagi semenjak kuliah aku dah gapernah mau ngajakin temen sholat lagi garagara tementemenku bilang dulu pas kuliah hybrid kita sholat soalnya gak enak sama kamudan ya emang semenjak kuliah offline gak ada sama sekali yang sholat	*apalagi, semenjak, kuliah, aku, dah, gapernah, mau, ngajakin, temen, sholat, lagi, garagara, tementemenku, bilang, dulu, pas, kuliah, hybrid, kita, sholat, soalnya, gak, enak, sama, kamudan, ya, emang, semenjak, kuliah, offline, gak, ada, sama, sekali, yang, sholat*

4.6 Stemming

Stemming is the process of mapping and parsing the various forms (variants) of a word into its basic word form. The stemming algorithm is developed based on the morphological rules of Indonesian, which categorize affixes into prefixes, infixes, suffixes, and confixes. As for this algorithm, it rearranges redundant words in a steaming process, or what is called recoding, and uses a dictionary of base words [15]. Here is an example of the word-stemming process.

Table 1. Example of Word Stemming

Sebelum	*Sesudah*
Trilliunan	*Trilliun*
Serendah	*Rendah*
Sebelum	*Belum*
Diberikan	*Beri*
Secukupnya	*Cukup*
Dilaksanakan	*Laksanakan*

4.7 Weighting Word

The flow at this stage begins with the process of handling data imbalance and TF-IDF weighting. The process of handling data imbalance, word vector, and TF-IDF weighting so that the resulting dataset is good for the classification modeling stage. The following formula is used in determining the Term Frequency-Inverse Document Frequency (TF-IDF) weight:

$$idf_t = log \frac{td}{df} \quad (1)$$

$$W_{t,d} = tf_{t,d} \times idf_t \quad (2)$$

After obtaining the labeled dataset, 2 data divisions are carried out, namely training data and testing data with a percentage ratio of 80:20 of the entire dataset that has been labeled. After dividing the training data and testing data, the training data will be used to build KNN and SVM classification models which will then produce a sentiment model.

The sentiment model that has been generated in the previous process will be used with testing data to test the classification model that has been created, followed by the classification process by calculating the probability of words for each class to produce data predictions.

Table 6. Number of Word Occurrences in Documents Using TF-IDF

Word	Attribution	Total	Documents	negative	neutral	positive
kuliah	*kuliah*	1727	1352	370	629	728
luring	*luring*	1508	1215	261	695	552
daring	*daring*	417	360	139	2	276
kampus	*kampus*	180	153	30	64	86
mahasiswa	*mahasiswa*	147	128	27	57	63

The data owned by 1564 tweet data is 660 data with a positive value, 312 data with a negative value, and 592 data with a neutral value. In sentiment analysis, the balance of data between sentiment categories (positive, negative, and neutral) has a very important role in maintaining the accuracy and reliability of the machine learning model used.

4.8 Balancing Data

The next step is to balance the imbalance data using the SMOTE technique. SMOTE (Synthetic Minority Oversampling Technique) is an oversampling method used in machine learning to handle class imbalance in datasets, especially in classification problems where the minority class has a low frequency compared to the majority class.

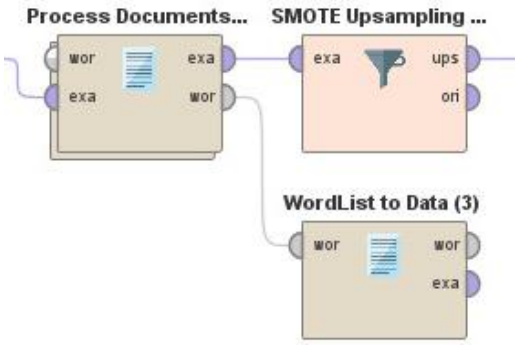

Figure 7. Data Balancing Operator

The SMOTE process involves creating a synthetic sample of the minority class by merging existing minority instances. The following are the results after balancing the data:

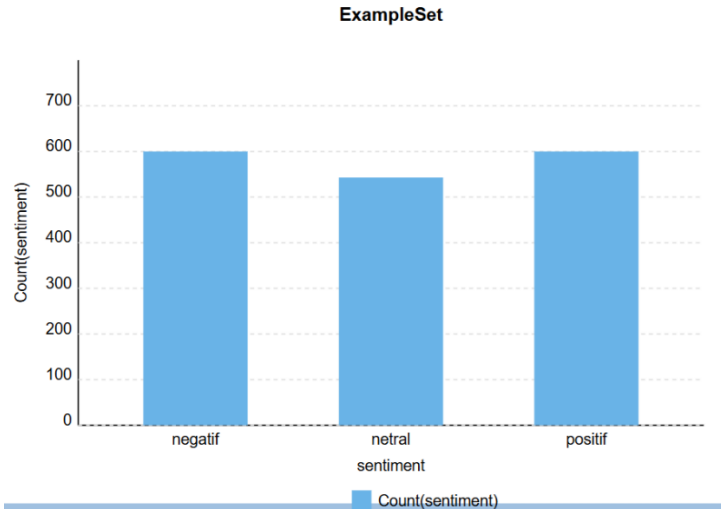

Figure 8. Number of sentiment datasets after SMOTE

4.9 Classification of Sentiment Analysis

After data pre-processing, cleaning, word weighting, and data balancing are done, the next stage is the sentiment analysis classification stage. This stage is the stage to provide training to implement various algorithms. In this research, two different classification operators will be used as defined earlier, namely K-NN and SVM.

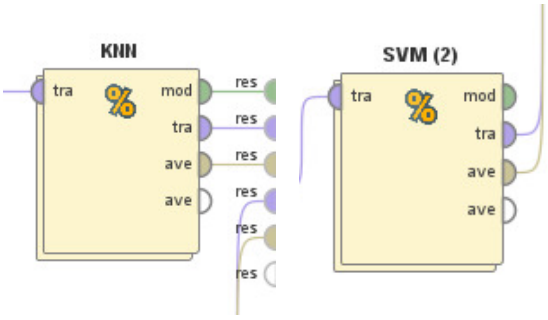

Figure 9. Operator classification

Moreover, the two classification operators will perform a split validation process accompanied by testing the model that has been trained along with testing the test data. The parameters used are split relative, split ratio: 0.9, and sampling type: automatic. The following are the results of the split validation that has been carried out in Table 6.

Table 2. Results of Split Validation of 2 models against test data

Data Train	Data Testing	SVM	KNN
90%	10%	63.39%	38.25%
80%	20%	61.58%	37.06%
70%	30%	61.23%	38.22%
60%	40%	59.38%	38.72%

4.10 Sentiment Analysis Evaluation

Classification evaluation aims to get an idea of how well the method used by the system performs in classifying the dataset. The classification performance test uses a confusion matrix. A confusion matrix is a matrix table that describes the performance of the classification model using testing data whose values are known.

Evaluation of the results is done with three parameters, namely, precision, accuracy, and recall, from a calculation that has been done [16]. Accuracy (A) is the number of correctly classified documents, both True Positive and True Negative. Calculating the accuracy value can use Equation 3.

$$A = \frac{TP+TN}{(TP+FP+TN+FN)} \times 100\% \quad (3)$$

Precision (P) is how much the processing result is relevant to the information you are looking for. In other words, precision is the True Positive classification, and all data is predicted as a positive class. Calculating the precision value can use Equation 4.

$$P = \frac{TP}{(TP+FP)} \times 100\% \quad (4)$$

Recall (R) is how many relevant documents in the collection are generated by the system. In other words, recall is the number of documents that have a True Positive classification out of all truly positive documents (including False Negative). Calculating the recall value can use Equation 4:

$$P = \frac{TP}{TP+FP} \times 100\% \quad (5)$$

Variables such as TN, TP, FN, and FP are derived from the confusion matrix. TN stands for True Negative, negative data that is classified as negative. True Positive stands for TP, which is positive data that will be classified as positive. False Negative stands for FN, which is positive data that will then be classified as negative data. FP stands for False Positive, negative data that is classified as positive [16]. For a more detailed explanation, see Table 7.

Table 3. Confussion Matrix

Classification		Predicted Class	
		Yes	**No**
Actual Class	Yes	True Positive (TP)	False Negative (FN)
	No	False Positive (FP)	True Negative (TN)

From the processing carried out on Rapidminer software, the results of the confusion matrix for the K-NN and SVM algorithms obtained on the dataset are as follows Table 8.

Table 4. Confusion Matrix of Each Algorithm

Methods	TP	FP	TN	FN
K-NN	1	0	63	61
SVM	55	19	42	2

The results of the average value of precision using Rapidminer software are shown in Table 9.

Table 9. Accuracy, Precision, and Recall

Method	Accuracy	Precision	Recall
K-NN	38.25%	54.27%	37,43%
SVM	63.39%	50.83%	62.38%

These results show that the accuracy of K-NN and SVM is 38.25% and 63.39%. The results for precision are 54.27% and 50.83%. While the results for Recall are 37.43% and 62.38%. So it can be concluded that the Support Vector Machine classifier is the best classifier to use with social media datasets because it provides more accurate and precise predictions.

5. Conclusion

Based on research on the comparison of the K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) methods in analyzing public sentiment towards offline lectures after the New Normal policy, it was found that the SVM method showed better performance than KNN. In testing several split data scenarios with a 90:10 ratio, the accuracy produced by SVM reached 63.39%, while KNN only reached 38.25%. In addition, the evaluation results show that SVM excels in Recall and F1-Score values compared to KNN.

This research also reveals that the SVM method is more effective in analyzing public sentiment towards offline lectures. With a higher accuracy rate, this method can classify comments better. In the analysis process, positive comments were found to be more dominant than negative comments, indicating that people's views tend to support the implementation of offline lectures after the New Normal policy.

Overall, the results of this study confirm that the SVM method is a superior choice over KNN for sentiment analysis in this context. With its advantages in accuracy and classification ability, this method can be used as an effective approach to understanding public opinion, especially regarding education policies during the new normal adaptation period.

Acknowledgments: Thanks to the Department of Information Systems, Jambi University, Indonesia for helping in the completion of this article.

Author contributions: The authors are responsible for building Conceptualization, Methodology, analysis, investigation, data curation, writing—original draft preparation, writing—review and editing, visualization, supervision of project administration, funding acquisition, and have read and agreed to the published version of the manuscript.

Funding: The study was conducted without any financial support from external sources.

Availability of data and Materials: All data are available from the authors.

Conflicts of Interest: The authors declare no conflict of interest.
Additional Information: No Additional Information from the authors.

References

[1] A. N. Yumang, J. M. Baguisi, B. R. S. Buenaventura and C. C. Paglinawan, "Detection of Black Sigatoka Disease on Banana Leaves Using ShuffleNet V2 CNN Architecture in Comparison to SVM and KNN Techniques," 2023 15th International Conference on Computer and Automation Engineering (ICCAE), Sydney, Australia, 2023, pp. 281-286, doi: 10.1109/ICCAE56788.2023.10111367.

[2] Touati., S. et. al. 2025.Performance analysis of steel W18CR4V grinding using RSM, DNN-GA, KNN, LM, DT, SVM models, and optimization via desirability function and MOGWO. Heliyon. Volume 11, Issue 4, 28 February 2025, e42640. doi.10.1016/j.heliyon.2025.e42640

[3] Yücesoy, Ergün. 2024. Gender Recognition from Speech Signal Using CNN, KNN, SVM and RF. Procedia Computer Science. Volume 235, 2024, Pages 2251-2257. doi. 10.1016/j.procs.2024.04.213

[4] Saxena., N. et. al. 2024. Hybrid KNN-SVM machine learning approach for solar power forecasting. Environmental Challenges. Volume 14, January 2024, 100838. doi. 10.1016/j.envc.2024.100838

[5] Çakir., M. et. al. 2023. Accuracy assessment of RFerns, NB, SVM, and kNN machine learning classifiers in aquaculture. Journal of King Saud University - Science. Volume 35, Issue 6, August 2023, 102754. doi. 10.1016/j.jksus.2023.102754

[6] S. Briouza, H. Gritli, N. Khraief, S. Belghith and D. Singh, "EMG Signal Classification for Human Hand Rehabilitation via Two Machine Learning Techniques: kNN and SVM," 2022 5th International Conference on Advanced Systems and Emergent Technologies (IC_ASET), Hammamet, Tunisia, 2022, pp. 412-417, doi: 10.1109/IC_ASET53395.2022.9765856.

[7] Q. D. P. Bayu, S. Suyanto and A. Arifianto, "Hierarchical SVM-kNN to Classify Music Emotion," 2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), Yogyakarta, Indonesia, 2019, pp. 5-10, doi: 10.1109/ISRITI48646.2019.9034651.

[8] V. Wasule and P. Sonar, "Classification of brain MRI using SVM and KNN classifier," 2017 Third International Conference on Sensing, Signal Processing and Security (ICSSS), Chennai, India, 2017, pp. 218-223, doi: 10.1109/SSPS.2017.8071594.

[9] C. Zhu, "Learning and Application of Different Machine Learning Methods (KNN, SVM, Decision Tree) in Different Datasets," 2024 5th International Conference on Machine Learning and Computer Application (ICMLCA), Hangzhou, China, 2024, pp. 1-5, doi: 10.1109/ICMLCA63499.2024.10754404.

[10] A. Zafar and A. Iqbal, "Machine Reading of Arabic Manuscripts using KNN and SVM Classifiers," 2020 7th International Conference on Computing for Sustainable Global Development (INDIACom), New Delhi, India, 2020, pp. 83-87, doi: 10.23919/INDIACom49435.2020.9083696.

[11] K. V. K. G, H. Shanmugasundaram, A. E, R. M, N. C and B. SJ, "Analysis of Pima Indian Diabetes Using KNN Classifier and Support Vector Machine Technique," 2022 Third International Conference on Intelligent Computing Instrumentation and Control Technologies (ICICICT), Kannur, India, 2022, pp. 1376-1380, doi: 10.1109/ICICICT54557.2022.9917992.

[12] E. Altayef, F. Anayi, M. Packianather, Y. Benmahamed and O. Kherif, "Detection and Classification of Lamination Faults in a 15 kVA Three-Phase Transformer Core Using SVM, KNN and DT Algorithms," in IEEE Access, vol. 10, pp. 50925-50932, 2022, doi: 10.1109/ACCESS.2022.3174359.

[13] R. Bathija, V. Bajaj, C. Megnani, J. Sawara and S. Mirchandani, "Revolutionizing Recruitment: A Comparative Study Of KNN, Weighted KNN, and SVM - KNN for Resume Screening," 2023 8th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2023, pp. 834-840, doi: 10.1109/ICCES57224.2023.10192665.

[14] I. F. Assalam, I. C. Gunadin and Syafaruddin, "Application of Support Vector Machine (SVM) Method for Photovoltaic Condition Classification Based on Characteristic Curve Indicators," 2023 International Conference on Artificial Intelligence Robotics, Signal and Image Processing (AIRoSIP), Yogyakarta, Indonesia, 2023, pp. 188-193, doi: 10.1109/AIRoSIP58759.2023.10873920.

[15] R. A. Dharmmesta, I. G. P. Jaya, A. Rizal and Istiqomah, "Classification of Foot Kicks in Taekwondo Using SVM (Support Vector Machine) and KNN (K-Nearest Neighbors) Algorithms," 2022 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT), BALI, Indonesia, 2022, pp. 36-41, doi: 10.1109/IAICT55358.2022.9887475.

[16] D. Hao and D. Chai, "Application of SVM-KNN intelligent classification prediction model in IT Vocational Education," 2021 5th Asian Conference on Artificial Intelligence Technology (ACAIT), Haikou, China, 2021, pp. 312-315, doi: 10.1109/ACAIT53529.2021.9731162.

[17] A. A. Bharate and M. S. Shirdhonkar, "Classification of Grape Leaves using KNN and SVM Classifiers," 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2020, pp. 745-749, doi: 10.1109/ICCMC48092.2020.ICCMC-000139.

[18] A. G. Karegowda, V. D and Y. Harshalatha, "Coconut Disease Detection using Traditional ML Classifiers with GLCM and SIFT Features," 2024 8th International Conference on Computational System and Information Technology for Sustainable Solutions (CSITSS), Bengaluru, India, 2024, pp. 1-6, doi: 10.1109/CSITSS64042.2024.10816784.

[19] Sotiropoulou, K. F. et. al. 2024. A hybrid AHP-PROMETHEE II onshore wind farms multicriteria suitability analysis using kNN and SVM regression models in northeastern Greece. Renewable Energy. Volume 221, February 2024, 119795. doi. 10.1016/j.renene.2023.119795

[20] Ghosh., S. et.al. 2022. SVM and KNN Based CNN Architectures for Plant Classification. Computers, Materials and Continua Computers, Materials and Continua. Volume 71, Issue 3, 13 January 2022, Pages 4257-4274. doi.10.32604/cmc.2022.023414