Scientific Journal of Informatics
Vol. 11, No. 1, Feb 2024
p-ISSN 2407-7658 http://journal.unnes.ac.id/nju/index.php/sji e-ISSN 2460-0040

Comparative Study of Machine Learning Algorithms for Performing Ham or Spam Classification in SMS

Erna Zuni Astuti[1], Christy Atika Sari[2*], Eko Hari Rachmawanto[3], Rabei Raad Ali[4]
[1, 2, 3]Departmen of Informatics Engineering, Universitas Dian Nuswantoro, Indonesia
[4]Department of Computer Engineering Technology, Northern Technical University, Iraq

Abstract.
Purpose: Fraud is rampant in the current era, especially in the era of technology where there is now easy access to a lot of information. Therefore, everyone needs to be able to sort out whether the information received is the right information or information that is fraudulent. In this research, the process of classifying messages including ham or spam has been carried out. The purpose of this research is to be able to build a model that can help classify messages. The purpose of this research is also to determine which machine learning method can accurately and efficiently perform the ham or spam classification process on messages.
Methods: In this research, the ham or spam classification process has been using machine learning methods. The machine learning methods used are the classification process with Random Forest, Logistic Regression, Support Vector Classification, Gradient Boosting, and XGBoost Classifier algorithms.
Results: The results obtained after testing in this study are the classification process using the Random Forest algorithm getting an accuracy of 97.28%, Logistic Regression getting an accuracy of 94.67%, with Support Vector Classification getting an accuracy of 97.93%, and using XGBoost Classifier getting an accuracy of 96.47%. The best precision value obtained in this study is 98% when using the random forest algorithm. The best recall value is 94% when using the SVC algorithm. While the best f1-score value is 95% when using the SVC algorithm.
Novelty: This research has been compared with several algorithms. In previous research, it is still very rarely done using XGBoost to classify the ham or spam in messages. We focus on giving brief information based con comparison algorithm and show the best algorithm to classify classify the ham or spam in messages. And for the novelty that exists from this research, the machine learning model built gets better accuracy when compared to previous research.

Keywords: Random forest, Logistic regression, Support vector classification, XGBoost classifier, Machine learning
Received September 2023 / **Revised** February 2024 / **Accepted** February 2024

INTRODUCTION

Now, technological advances are growing rapidly, especially in the telecommunications sector [1], [2]. However, because of this, can also lead to digital crime [3]. Many frauds are carried out digitally, one of the media is using SMS [4]. This happens because SMS is easy to use and does not require expensive costs to send messages [5]. As well as the fact that SMS does not require the internet [6], so sending messages can be done at any time and more easily. Salman et. al [7] argued, in 2021 in the United States, at least an estimated 86 million USD was lost due to fraud via SMS. The fraud can also take the form of spam which is very disturbing to SMS users [8] because it usually has less important information [9]. Spam emails are unsolicited, often irrelevant or malicious messages sent in bulk to a large number of recipients without their consent [10]. SMS spam may be exploited to spread hate speech and criminal activity, or to disseminate commercial information, commonly known as spam [11]. Machine learning is part of the Artificial Intelligence (AI) process [12]. Machine Learning is also included in the computer science family where in ML, computers can learn independently without having to be programmed first [13]. It can also be interpreted that the machine learning model can carry out the learning process automatically when it has been given an algorithm that is used to handle the given task [14]. So that it can perform intelligent data analysis automatically [15]. This can happen because the algorithm in machine learning uses a

*Corresponding author.
Email addresses: erna.zuni.astuti@dsn.dinus.ac.id (Astuti), christy.atika.sari@dsn.dinus.ac.id (Sari)*, eko.hari@dsn.dinus.ac.id (Rachmawanto), rabei@ntu.edu.iq (Ali)
DOI: 10.15294/sji.v11i1.47364

mathematical approach to learn patterns from data [16] so that it can handle data more efficiently [17]. Therefore, the model built can properly update and also adjust to the actions that have been taken [18]. That means the results of the analysis and learning of the model can be used for prediction, estimation, or classification [19].

Random forest is a classification technique that is processed by combining several models [20] decision tree to make a forest [21] that can be used for data classification. So that the use of many trees can produce better accuracy [22]. Logistic Regression is a type of supervised classification [23], so it requires a label as a target. In its implementation, logistic regression has excellent performance in predicting discrete probabilities (only having 2 classes) [24]. This can be done because, in logistic regression, the probability value of an event is used as a logistic function [25]. Support Vector Classification is a classification algorithm that has the concept of Vapnik's statistical theory [26]. In the process, SVC is the same as SVM, that is minimizing the distance between the Support Vector and the data by using the maximum margin cost [27]. XGBoost is an algorithm that is the realization of Gradient Boosting Decision Tree [28]. This algorithm is also included in the ensemble algorithm that combines several decision tree models [29].

In the process, XGBoost is used to improve the decision tree so that the tree model built does not experience overfitting [30]. This research discussed about classification process of ham or spam in SMS. The purpose of this research is to build a classification model that can perform the ham or spam classification process from SMS text. Later is expected that the model built can help users determine whether the message received is a ham or spam message. This research used 4 machine learning algorithms for the classification process, namely Random Forest, Logistic Regression, Support Vector Classification, and also XGBoost. The purpose of using these 4 algorithms is to later compare and find which algorithm is the best in classifying ham or spam. The purpose of using Random forest is because this algorithm is an ensemble learning model that combines several decision trees, so it is expected to have good performance for classification. The purpose of using logistic regression is because it is an excellent algorithm for classification which only has 2 main classes. The purpose of using Support Vector Classification is because this algorithm runs like SVM which uses support vectors and is good at classifying data. The purpose of using the XGBoost algorithm is because this algorithm is an ensemble learning algorithm but has an improvement process from each tree model built, so it is expected to provide good performance for data classification.

Research that had conducted by Jakins et. Al [31], discusses building AI models that can predict diseases using random forest classifiers and naïve bayes. The purpose of this research is to build a model that can accurately predict diabetes, coronary heart disease, and breast cancer. The results obtained after this research are for diabetes prediction, the naïve Bayes algorithm gets better test accuracy of 74.46%. For heart disease prediction, the random forest algorithm gets better testing accuracy of 83.85 and for breast cancer prediction, the random forest algorithm gets better testing accuracy of 92.40. Research conducted by Shah et. al [23] in 2020 discussed about comparative study of the Logistic regression algorithm, random forest, and K-Nearest Neighbors to be able to classify text. The purpose of this research is to find which algorithm is more effective for classifying text. The results that would obtained from this study are that after testing the logistic regression algorithm gets the best accuracy for classification which is 97%.

Research conducted by Barman et al [32] discusses the classification of soil texture using multi-class SVM. The purpose of this research is to process soil images to build a soil classification system that can be used by rural farmers at a low cost. The results after doing this research are using 3 soil classes, getting a test accuracy of 95.72%, and using 12 soil classes getting a test accuracy of 91.37%. Research conducted by Qi [33] in 2020 discussed the classification of theft based on text. The purpose of the study is to determine the performance of the XGBoost classifier when compared to other machine learning algorithms to be able to perform classification. The results obtained from this study are that after several adjustments, the XGBoost algorithm gets the best results, namely with a precision value of 96.8%, a recall value of 96.4%, and an f1-score value of 96.6%. Previous research conducted by Kudupudi in 2021 [34] also discussed the process of classifying spam or spam messages using the same dataset. The study, using the logistic regression algorithm got a research accuracy of 96%.

METHODS

In this research, we used the SMS spam dataset obtained from the kaggle.com website. The data to be used consists of 2 main classes, namely ham and spam, which have a total of 5574 data. Which where the data consists of 4825 ham data and 747 spam data. This research is focused on SMS for the classification of ham or spam messages using machine learning, namely because SMS data is available in large quantities, provides long-term validity for algorithms to practice classification, makes it easy to adapt to other message types, and is still relevant to most users in various regions. Figure 1 shows the visualization from the dataset used in this study. From the total of 5574 data, divided into training data and testing data. With a percentage of 70% train data and 30% test data. This dataset can be used as an initial foundation for developing and validating ham or spam classification methods with machine learning. In addition, although the data is from 2012, the basic concepts in spam message classification are still relevant and can be adapted to current trends. Therefore, while this dataset may not reflect today's real-world SMS completely, the results of this study can provide useful preliminary insights for the development of more advanced methods that can be applied to larger datasets and more actual contexts in the future. To conduct training and also testing processes, the data must first be processed. The processing that is done in this research is converting text data into vectors, it can be more easily processed by the machine learning algorithms. In this research, a statistical approach that can be converted text into a vector would be used, namely TF-IDF or Term Frecuency/Inverse Document Frecuency [35]. Because in TF-IDF the statistical method used converts from text into a number based on the importance of the word in documents and Scopus [36]. The weight of the word is also obtained when the analysis process is carried out [37]. The formula to find TF (how often the word appears) is given in equation 1, while to find IDF (how unique the word that appears) is given in equation 2 and for weight search (TF-IDF value) is given in equation 3. Although TF-IDF has limitations in capturing semantics, it is effective in identifying keywords that distinguish ham messages from spam, especially with limited data. Whereas semantic methods such as Word2Vec or GLOVE require larger data and strong computational resources. The choice of method depends on resources and research objectives, so the use of TF-IDF provides adequate results with limited resources, with consideration of alternative methods according to the research context.

	label	text
0	ham	Go until jurong point, crazy.. Available only ...
1	ham	Ok lar... Joking wif u oni...
2	spam	Free entry in 2 a wkly comp to win FA Cup fina...
3	ham	U dun say so early hor... U c already then say...
4	ham	Nah I don't think he goes to usf, he lives aro...

Figure 1. Dataset that used

$$TF(W, Doc) = \frac{\sum(Word\ appears\ in\ Doc)}{\sum(All\ words\ in\ Doc)} \quad (1)$$

$$IDF(W, Corp) = \log\left(\frac{N}{(1 + DF(W, Corp))}\right) \quad (2)$$

$$TFIDF(W, Doc, Corp) = TF(W, Doc) * IDF(W, Corp) \quad (3)$$

This paper may not have directly considered the relevance of SMS messages or presented the most advanced method. However, the goal was to develop a solid foundation in the classification of spam or spam messages based on text characteristics. The use of traditional algorithms in this research is a reasonable step to understand the basics of SMS message classification before considering more complex methods. SMS messages may vary in urgency and importance, but the results of this study can be a useful basis for further development, including the assessment of message urgency levels. Thus, although this study may have limitations, it provides a solid foundation for further research in the domain of SMS message classification. After the text transformation process into a vector, the data can be used for learning and testing machine learning models. This research has used machine learning models with Random Forest, Logistic Regression, Support Vector Classification, and XGBoost classifier algorithms. Random forest is an algorithm that in the process of combining several models [20] decision tree to form a forest [21] where in the end the classification process can be carried out from the forest that has been built. So it can also be

interpreted that random forest is an algorithm that includes ensemble learning or an algorithm that combines several algorithms to be able to perform a better classification process. The stages of the random forest process can be described below.

Define N Tree

(1) Select randomly X feature from the data

(2) For each I in X

(1) $Ent\ (Z) = -\sum_{x=1}^{n} P_x\ log_2(P_x)$

(2) $Ent\ (Z,I) = \sum_{m \in I} P(m)E(m)$

(3) $InformationGain\ (C,Z) = Ent\ (C) - Ent(Z,I)$

(4) Select node X which has the highest information gain

(5) Split node into sub-node

Repeat steps 1 to 5 until construct a tree and reach the minimum number of samples that are required

(3) Repeat steps 1 to 2 N times until building a forest of N trees

Logistic Regression is a type of supervised classification [23] which in its implementation, has excellent performance in predicting discrete probabilities [24]. This can be done because, in logistic regression, the probability value of an event is used as a logistic function [25]. So by using the logistic function, the resulting output is 0 or 1. Therefore, the logistic regression process is very good for classifying binary classes. For logistic regression, the pseudocode is given below.

#Define param

Define w (weight), b (bias), $\propto$ (learn rate), iteration

Logistic Function

(1) $\sigma(x) = \frac{1}{(1+e^{-z})}$

iteration training start

(2) Repeat for iteration

Calculate combination

(1) x = w * feature + b

apply function

(2) prob = $\sigma(x)$

calculate the loss function

(3) loss = $-(target * \log(prob) + (1 - target) * \log(1 - prob))$

Calculate gradients

(4) dw = $\frac{1}{total\ data} * \sum\big((prob - target) * feature\big)$

(5) db = $\frac{1}{total\ data} * \sum(prob - target)$

Update weight

(6) w = w * α * dw

Update bias

(7) b = b * α * db

Prediction

(3) if prob > 0.5 then 1 else 0

Support Vector Classification is a classification algorithm that has the concept of Vapnik's statistical theory [26]. In the process, SVC is the same as SVM, namely by minimizing the distance between the Support Vector and the sample by using the maximum margin cost [27]. Because it has the same process as SVM, SVC also requires a hyperplane to make predictions. For the pseudocode of SVC is given below.

Define param

Define w (weight), b (bias), $\propto$ (learn rate), Z (regularization param)

Iteration training start

(1) Repeat until convergence (i)

Calculate margin

(1) Margin = m(i) * (w * n(i) * b)

(2) if Margin < 1 then
update w
(1) w = $w + \alpha * (m(i) * n(i) - 2 * Z * w)$
update b
(2) b = $b + \alpha * m(i)$
#
(3) else
update w
(1) w = $w + \alpha * (-2 * Z * w)$
update b
(2) b = b
predict new x
(2) f(x) = w * x * b
prediction
(3) if f(x) > 0 then 1 else 0

XGBoost is an algorithm that is a realization of Gradient Boosting Decision Tree [28]. This algorithm is also included in the ensemble algorithm that combines several decision tree models [29]. In the process, XGBoost is used to improve the decision tree so that the tree model built does not experience overfitting [30]. Therefore, in this process, it is expected that the model built can make good and optimal predictions. For XGBoost pseudocode is given below.

define param and data
Define model = [], num boost (iteration), x, y
training iteration start
(1) Repeat for iteration
gradient loss function from the current ensemble
(1) gradientMin = $-\nabla L(y, \text{predict1})$
train for decision tree
(2) base = train decision tree classifier (x,y)
predict the probability distribution of gradientMin
(3) base = predict(GradientMin)
Add train model to the ensemble
(4) model.append(base)
update ensemble prediction
(5) predict1 = $predict1 + learning\ rate * softmax(base\_predict)$

In the model that had been carried out for the training process, the model has been tested. After the testing process, the performance of the model can be calculated. In this study, the process of calculating the performance of the model would use a confusion matrix where in the matrix, which model guesses are correct or wrong during the testing process. The values from the confusion matrix are positive true, positive false, negative true, and negative false. From these values, model performance can be calculated by namely accuracy, precision, recall, and also f1-score. The calculation formulas for precision, recall, and f1-score are given in equations 4, 5, and 6.

$$Precision = \frac{TruP}{(TruP + FalP)} \tag{4}$$

$$Recall = \frac{TruP}{TruP + FalN} \tag{5}$$

$$f1 - Score = 2 * \frac{(Prec * Rec)}{(Prec + Rec)} \tag{6}$$

Data has been split into 2 processes, namely the training process and the testing process. The training process aims to make the model that had been built with machine learning algorithms learn patterns from the data given. Meanwhile, the model testing process aims to make the model that has been trained to recognize the pattern can be tested so that the performance of the model is known to be able to carry out

the ham or spam classification process based on SMS text. The explanation of each stage is given below. The flow of the classification research method is given in Figure 2, as follows:

1. The first thing is to read the data for the classification process.
2. After the data reading process is carried out, then the transformation process from text to vector has been carried out. The purpose of this process is so that the data that has been read can later be used for the training process and also model testing.
3. After transforming the data to vectors, then divided into train data and test data. With a percentage of 70% train data and 30% test data.
4. Then, build a model that would be used for classification. The models that were built are models with Random Forest, Logistic Regression, Support Vector Classification, and XGBoost Classifier algorithms. The parameters used in each model are given in Table 1.
5. After building the model along with the parameters used, the training and testing process of the model has been carried out with the training and testing data that was previously divided.
6. Then, after completing testing on the model, the performance calculation of the model test results can be carried out. The performance calculation has been implemented using the value of the confusion matrix, namely accuracy, precision, recall, and f1-score.

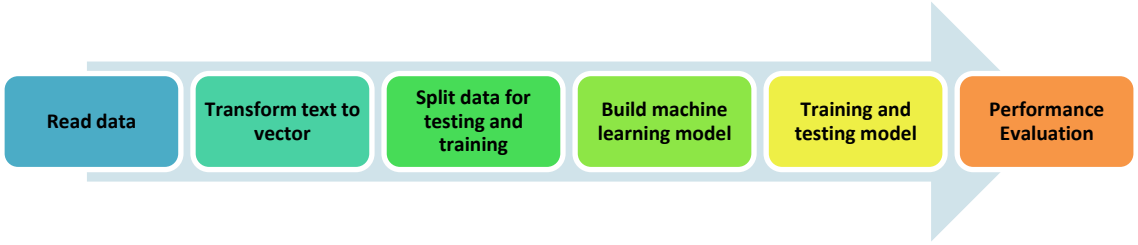

Figure 2. Workflow process for classification

Table 1. Parameter for Machine Learning Algorithm

Random Forest		Logistic Regression		Support Vector Classification		XGBoost Classifier	
Parameter	Value	Parameter	Value	Parameter	Value	Parameter	Value
n_estimators	500	penalty	L2	C	1.0	objective	Binary:logistic
criterion	Entropy	dual	False	kernel	sigmoid	n_estimators	100
max_depth	None	tol	0.0	degree	3	learning_rate	0.1
min_samples_split	2	C	1.0	gamma	1.0	max_depth	3
min_samples_leaf	1	fit_intercept	True	shrinking	True	random_state	42
min_weight_fraction_leaf	0.0	class_weight	None	probability	False	booster	GBtree
max_features	Auto	random_state	None	tol	0.001	gamma	0
max_leaf_nodes	None	solver	Sag	class_weight	Balanced	min_child_weight	1
min_impurity_decrease	0.0	max_iter	100	verbose	False	max_delta_step	0
bootstrap	True	multi_class	Auto	decision_function_shape	Ovr	subsample	1
oob_score	False	warm_start	False	random_state	None	colsample_bytree	1
n_jobs	-1	n_jobs	None			colsample_bylevel	1
random_state	42					colsample_bynode	1
						reg_alpha	0
						reg_lambda	1
						scale_pos_weight	1
						base_score	0.5
						missing	Np.nan

RESULTS AND DISCUSSIONS

In this research, the training and testing process has used Jupyter Notebook tools as an IDE to write and run the program. For system implementation, this research used the Python programming language. After completing data reading, data preprocessing, and model building, then the model training process can be carried out so that the model built can learn existing patterns from data that has previously been preprocessed. After training, the model that has been trained has been tested. The results of the test in the form of test accuracy are given in Table 2.

Table 2 shows the accuracy obtained after the testing process for ham or spam classification. As seen in Table 2, the best accuracy is obtained when the classification process using the Support Vector Classification algorithm is 97.93%. This shows that the Support Vector Classification model can accurately perform the ham or spam classification process. Meanwhile, the lowest accuracy value obtained is 94.67%, which is obtained when performing the classification process using the Logistic Regression algorithm. However, as seen in Table 2, all accuracies range more than 94%, which means that the classification process in each model can run well. At the time of testing, the value of the confusion matrix obtained can also be calculated. The confusion matrix value obtained is given in Figure 3. Figure 3 shows the value of the confusion matrix of the model test results. From this value, had been seen that the value of the correct positive value, the wrong positive, the correct negative, and also the wrong negative. These values can be used to calculate the precision, recall, and f1-score values. That value can used to see and calculate the performance of the model when performing the classification process. The results of the precision, recall, and f1-score values are given in Table 3.

Table 2. Accuracy comparison from a model that builds

Algorithm	Accuracy
Random Forest	97.28%
Logistic Regression	94.67%
Support Vector Classification	97.93%
XGBoost Classifier	96.47%

Table 3. Precision, Recall, F1-Score, and Support from the testing process

Algorithm	Avg Precision	Avg Recall	Avg F1-Score	Support
Random Forest	98%	90%	94%	1839
Logistic Regression	97%	80%	86%	1839
Support Vector Classification	97%	94%	95%	1839
XGBoost Classifier	96%	89%	92%	1839

Table 3 shows the results of precision, recall, and also the f1-score obtained after testing the model. Had is seen in Table 3 that the best average precision value is obtained when the classification process uses the Random Forest algorithm. This means that the random forest algorithm had good accuracy that able to carry out the classification process. As for the average of the best recall, namely when performing the classification process using Support Vector Classification. This shows that when using the Support Vector Classification model having good performance for predicting all classes. And for the average value of the best f1-score obtained when performing the classification process with Support Vector Classification, is 95%. That value shows that when the classification process uses Support Vector Classification, it gets a very good harmonic value between precision and recall. The support value for all algorithms is the same, which is 1839 because the support value is the value of the amount of data used for the testing process. Some research has been done on the ham or spam classification process. Table 4 gives some related research that discusses the ham or spam classification process. Table 4 shows research that discusses about ham or spam classification process. Table 4 also shows the comparison between the methods used in this research and previous research. Had been seen that the method used in this study was successful in improving the accuracy for the ham or spam classification task, this is evidenced by the accuracy obtained in this study has increased when compared to research that has been done before.

Table 4. Comparison from previous research

Journal	Algorithm or method	Accuracy
Annareddy [38]	CNN and RNN	CNN is 96.4% and RNN is 97,8%
Kudupudi [34]	Logistic Regression	96%
Alzahrani [39]	Logistic Regression, Naïve Bayes, SVC, Neural Network	LR is 94.26%, NB is 88.16%, SVC is 94.26% and NN is 97.67%
Shobana [40]	Multinominal Naïve Bayes with Passive Aggressive Algorithm	89%
Our Proposed Scheme	Random Forest, Logistic Regression, Support Vector Classification, XGBoost Classifier	RF is 97.28%, LR is 94.67%, SVC is 97.93% and XGB is 96.47

CONCLUSION

After testing and analyzing the ham or spam classification process using machine learning algorithms, the best accuracy result obtained is 97.93%, where the accuracy is obtained when performing the classification process using the Support Vector Classification algorithm. So from this research, it can be concluded that the Support Vector Classification algorithm is an algorithm that can properly and accurately perform the ham or spam classification process based on SMS text. Based on the results that have been obtained, it can be seen that the model built in this study gets better accuracy and there is an increase in accuracy from previous studies.

In future research, it is hoped that the classification process can use the neural network method so that later the performance of the neural network can be seen in the classification. For further research, it is also hoped that it can be added to be able to carry out the classification process using Indonesian news or text and can also use other processing methods such as word2vec or glove and be able to add more parameters that are used in random forest, logistic regression, support vector classification or XGBoost classifier algorithms so that later a more in-depth performance analysis can be carried out in the data classification process.

REFERENCES

[1] N. Duha, "Short Message Services (SMS) Fraud Against Mobile Telephone Provider Consumer Review From Law Number 8 Of 1999 Concerning Consumer Protection," *J. Law Sci.*, vol. 3, pp. 36–43, Jan. 2021, doi: 10.35335/jls.v3i1.1654.

[2] N. P. Damayanti, D. E. Prameswari, W. Puspita, and P. S. Sundari, "Classification of Hate Comments on Twitter Using a Combination of Logistic Regression and Support Vector Machine Algorithm," *J. Inf. Syst. Explor. Res.*, vol. 2, no. 1, Jan. 2024, doi: 10.52465/joiser.v2i1.229.

[3] Y. Zhang, Z. Wang, Z. Wang, and C. Liu, "A Robust and Adaptive Watermarking Technique for Relational Database," 2022, pp. 3–26. doi: 10.1007/978-981-16-9229-1_1.

[4] O. F. Cossa, N. Sousa, R. Goncalves, J. Martins, and F. Branco, "Prediction of bank frauds by SMS or voice, from cell phone data analysis: A Systematic Literature Review," in *2021 16th Iberian Conference on Information Systems and Technologies (CISTI)*, IEEE, Jun. 2021, pp. 1–8. doi: 10.23919/CISTI52073.2021.9476380.

[5] A. Wagstaff, E. van Doorslaer, and R. Burger, "SMS nudges as a tool to reduce tuberculosis treatment delay and pretreatment loss to follow-up. A randomized controlled trial," *PLoS One*, vol. 14, no. 6, p. e0218527, Jun. 2019, doi: 10.1371/journal.pone.0218527.

[6] H. R. Nafiisah and F. Z. Ruskanda, "Content-based Multiclass Classification on Indonesian SMS Messages," in *2022 International Symposium on Electronics and Smart Devices (ISESD)*, IEEE, Nov. 2022, pp. 1–6. doi: 10.1109/ISESD56103.2022.9980769.

[7] M. Salman, M. Ikram, and D. Kaafar, *An Empirical Analysis of SMS Scam Detection Systems*. 2022.

[8] H. H. Mansoor* and A. P. D. S. H. Shaker, "Using Classification Techniques to SMS Spam Filter," *Int. J. Innov. Technol. Explor. Eng.*, vol. 8, no. 12, pp. 1734–1739, Oct. 2019, doi: 10.35940/ijitee.L3206.1081219.

[9] V. V Sergeev, I. M. Gorbchenko, and V. V Safronov, "Comparative analysis of fraud detection systems by phone number," *J. Phys. Conf. Ser.*, vol. 1679, no. 5, p. 052003, Nov. 2020, doi: 10.1088/1742-6596/1679/5/052003.

[10] D. Budiman, Z. Zayyan, A. Mardiana, and A. A. Mahrani, "Email spam detection: a comparison of svm and naive bayes using bayesian optimization and grid search parameters," *J. Student Res. Explor.*, vol. 2, no. 1, pp. 53–64, Jan. 2024, doi: 10.52465/josre.v2i1.260.

[11] "Performance comparison of support vector machine and gaussian naive bayes classifier for youtube spam comment detection," *J. Soft Comput. Explor.*, vol. 2, no. 2, Sep. 2021, doi: 10.52465/joscex.v2i2.42.

[12] A. Dogan and D. Birant, "Machine learning and data mining in manufacturing," *Expert Syst. Appl.*, vol. 166, p. 114060, Mar. 2021, doi: 10.1016/j.eswa.2020.114060.

[13] Q. Bi, K. E. Goodman, J. Kaminsky, and J. Lessler, "What is Machine Learning? A Primer for the Epidemiologist," *Am. J. Epidemiol.*, Oct. 2019, doi: 10.1093/aje/kwz189.

[14] L. Zhang *et al.*, "A review of machine learning in building load prediction," *Appl. Energy*, vol. 285, p. 116452, Mar. 2021, doi: 10.1016/j.apenergy.2021.116452.

[15] I. H. Sarker, "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput. Sci.*, vol. 2, no. 3, p. 160, May 2021, doi: 10.1007/s42979-021-00592-x.

[16] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Med. Inform. Decis. Mak.*, vol. 19, no. 1, p. 281, Dec. 2019, doi: 10.1186/s12911-019-1004-8.

[17] B. Mahesh, *Machine Learning Algorithms -A Review*. 2019. doi: 10.21275/ART20203995.
[18] S. Kushwaha *et al.*, "Significant Applications of Machine Learning for COVID-19 Pandemic," *J. Ind. Integr. Manag.*, vol. 05, no. 04, pp. 453–479, Dec. 2020, doi: 10.1142/S2424862220500268.
[19] T. Meng, X. Jing, Z. Yan, and W. Pedrycz, "A survey on machine learning for data fusion," *Inf. Fusion*, vol. 57, pp. 115–129, May 2020, doi: 10.1016/j.inffus.2019.12.001.
[20] A. Parmar, R. Katariya, and V. Patel, "A Review on Random Forest: An Ensemble Classifier," 2019, pp. 758–763. doi: 10.1007/978-3-030-03146-6_86.
[21] A. Subudhi, M. Dash, and S. Sabut, "Automated segmentation and classification of brain stroke using expectation-maximization and random forest classifier," *Biocybern. Biomed. Eng.*, vol. 40, no. 1, pp. 277–289, Jan. 2020, doi: 10.1016/j.bbe.2019.04.004.
[22] M. Huljanah, Z. Rustam, S. Utama, and T. Siswantining, "Feature Selection using Random Forest Classifier for Predicting Prostate Cancer," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 546, no. 5, p. 052031, Jun. 2019, doi: 10.1088/1757-899X/546/5/052031.
[23] K. Shah, H. Patel, D. Sanghvi, and M. Shah, "A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification," *Augment. Hum. Res.*, vol. 5, no. 1, p. 12, Dec. 2020, doi: 10.1007/s41133-020-00032-0.
[24] N. M. Samsudin, C. F. binti Mohd Foozy, N. Alias, P. Shamala, N. F. Othman, and W. I. S. Wan Din, "Youtube spam detection framework using naïve bayes and logistic regression," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 14, no. 3, p. 1508, Jun. 2019, doi: 10.11591/ijeecs.v14.i3.pp1508-1517.
[25] K. V. Kumar and M. Ramamoorthy, "Machine Learning-based spam detection using Naïve Bayes Classifier in comparison with Logistic Regression for improving accuracy," *J. Pharm. Negat. Results*, vol. 13, no. SO4, Jan. 2022, doi: 10.47750/pnr.2022.13.S04.061.
[26] D. Xia, H. Tang, S. Sun, C. Tang, and B. Zhang, "Landslide Susceptibility Mapping Based on the Germinal Center Optimization Algorithm and Support Vector Classification," *Remote Sens.*, vol. 14, no. 11, p. 2707, Jun. 2022, doi: 10.3390/rs14112707.
[27] O. Rákos, S. Aradi, and T. Bécsi, "Lane Change Prediction Using Gaussian Classification, Support Vector Classification and Neural Network Classifiers," *Period. Polytech. Transp. Eng.*, vol. 48, no. 4, pp. 327–333, Jun. 2020, doi: 10.3311/PPtr.15849.
[28] R. Zhang, B. Li, and B. Jiao, "Application of XGboost Algorithm in Bearing Fault Diagnosis," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 490, p. 072062, Apr. 2019, doi: 10.1088/1757-899X/490/7/072062.
[29] I. L. Cherif and A. Kortebi, "On using eXtreme Gradient Boosting (XGBoost) Machine Learning algorithm for Home Network Traffic Classification," in *2019 Wireless Days (WD)*, IEEE, Apr. 2019, pp. 1–6. doi: 10.1109/WD.2019.8734193.
[30] S. Thongsuwan, S. Jaiyen, A. Padcharoen, and P. Agarwal, "ConvXGB: A new deep learning model for classification problems based on CNN and XGBoost," *Nucl. Eng. Technol.*, vol. 53, no. 2, pp. 522–531, Feb. 2021, doi: 10.1016/j.net.2020.04.008.
[31] V. Jackins, S. Vimal, M. Kaliappan, and M. Y. Lee, "AI-based smart prediction of clinical disease using random forest classifier and Naive Bayes," *J. Supercomput.*, vol. 77, no. 5, pp. 5198–5219, May 2021, doi: 10.1007/s11227-020-03481-x.
[32] U. Barman and R. D. Choudhury, "Soil texture classification using multi class support vector machine," *Inf. Process. Agric.*, vol. 7, no. 2, pp. 318–332, Jun. 2020, doi: 10.1016/j.inpa.2019.08.001.
[33] Z. Qi, "The Text Classification of Theft Crime Based on TF-IDF and XGBoost Model," in *2020 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA)*, IEEE, Jun. 2020, pp. 1241–1246. doi: 10.1109/ICAICA50127.2020.9182555.
[34] N. Kudupudi and S. Nair, "Spam Message Detection Using Logistic Regression," 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:259939313
[35] J. Wang, W. Xu, W. Yan, and C. Li, "Text Similarity Calculation Method Based on Hybrid Model of LDA and TF-IDF," in *Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence*, New York, NY, USA: ACM, Dec. 2019, pp. 1–8. doi: 10.1145/3374587.3374590.
[36] R. S. Patil and S. R. Kolhe, "Supervised classifiers with TF-IDF features for sentiment analysis of Marathi tweets," *Soc. Netw. Anal. Min.*, vol. 12, no. 1, p. 51, Dec. 2022, doi: 10.1007/s13278-022-00877-w.
[37] V. Kumar and B. Subba, "A TfidfVectorizer and SVM based sentiment analysis framework for text data corpus," in *2020 National Conference on Communications (NCC)*, IEEE, Feb. 2020, pp. 1–6.

doi: 10.1109/NCC48643.2020.9056085.
[38] S. Annareddy and S. Tammina, "A Comparative Study of Deep Learning Methods for Spam Detection," in *2019 Third International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC)*, IEEE, Dec. 2019, pp. 66–72. doi: 10.1109/I-SMAC47947.2019.9032627.
[39] A. Alzahrani and D. B. Rawat, "Comparative Study of Machine Learning Algorithms for SMS Spam Detection," in *2019 SoutheastCon*, IEEE, Apr. 2019, pp. 1–6. doi: 10.1109/SoutheastCon42311.2019.9020530.
[40] J. Shobana and D. Kanchana, "An Efficient Spam SMS Analysis Model based on Multinomial Naïve Bayes model Using Passive Aggressive Algorithm," *J. Phys. Conf. Ser.*, vol. 2007, no. 1, p. 012047, Aug. 2021, doi: 10.1088/1742-6596/2007/1/012047.