# AN INNOVATIVE METHOD TO ENHANCE THE ACCURACY IN CLASSIFICATION OF SPAM DETECTION FOR YOUTUBE COMMENTS WITH USING DECISION TREE OVER K–NEAREST NEIGHBOR

**S. Venkata Manoj Kumar Reddy[1], K. Malathi[2 *]**

**Article History: Received:** 12.12.2022 **Revised:** 29.01.2023 **Accepted:** 15.03.2023

**Abstract**

**Aim:** The objective of this research paper is to detect the spam comments on YouTube videos with an enhanced accuracy rate by using Novel Decision Tree (D-Tree) in comparison with K-Nearest Neighbor (KNN) Classifier.
**Materials & Methods:** The data set in this paper utilizes UCI machine learning repositories. The sample size of predicting the spam comments on YouTube videos with enhanced accuracy rate was sample 80 (Group 1=40 and Group 2 =40) and calculation is performed utilizing G-power 0.8 with alpha and beta qualities are 0.05, 0.2 with a confidence interval at 95%. Predicting the spam comments on YouTube videos with enhanced accuracy rate is performed by Novel Decision Tree (D-Tree) whereas the number of samples (N=10) and K-Nearest Neighbor (KNN) were the number of samples (N=10).
**Results:** The Novel Decision Tree (D-Tree) classifier has 94.47% higher accuracy rates when compared to the accuracy rate of K-Nearest Neighbor (KNN) is 86.91%. The study has a significance value of $p<0.05$ i.e. $p=0.0291$.
**Conclusion:** The Novel Decision Tree (D-Tree) provides the better outcomes in accuracy rate when compared to K-Nearest Neighbor (KNN) for detecting the spam comments on YouTube videos with enhanced accuracy rate.

**Keywords:** Spam Detection, Novel Decision Tree Classifier, K-Nearest Neighbor Classifier, Accuracy Rate, Youtube Spam, Machine Learning.

[1]Research Scholar, Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India. Pincode:602105.
[2*]Department of Computer Science and Engineering, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamil Nadu, India. pincode: 602105.

## 1. Introduction

The aim of the research paper is to detect the spam comments on YouTube videos with an enhanced accuracy rate. YouTube is one of the biggest sites for users to get information on the Internet (Schultes, Dorner, and Lehner 2013). Because of that, many spammers will trick the YouTube user by spamming the YouTube comments. According to Hamou (Hamou and Amine 2013), spam is now a trend attack and YouTube defines spam as inappropriate comments, such as abuse or trolling, and also people trying to sell things. Ham can be defined as "good comments" for YouTube free from spam comments. Spam can be categorized as dangerous because spam has the potential of cyber security threat for end users. The spammer used this opportunity to spread malware through comment fields, which will exploit vulnerabilities in the user's machines. Another intention includes seizing money transactions and hijacking credit card and banking information. Besides, spammers tend to ruin the content of web pages. This action will lead visitors to be annoyed overall with the posted content (Alsaleh et al. 2015). This paper proposes a new approach by comparing the analysis results using an Innovative Decision Tree (D-Tree) and the K-Nearest Neighbor classification. The experimental results show that the proposed Decision Tree (D-Tree) model gives a superior performance than the existing K-Nearest Neighbor (KNN) model (Aiyar and Shetty 2018).

There are several studies to detect YouTube Spam such as proposed to classify the YouTube comment as Spam and Ham by using machine learning algorithms. IEEE Explore published 105 research papers, and Google Scholar found 97 articles. The paper by (Jin et al. 2011) focuses on clearing out spam videos and at the same time also identifies the spammers responsible. The classification is done with ID3, K Nearest Neighbors, and Innovative Decision Tree (D-Tree) algorithms. Distinguishing feature selector and Gini Index feature selection methods have been reported to provide the best classification results in spam filtering tasks on YouTube (Alberto, Lochter, and Almeida 2015b). The authors recommend Decision Tree (D-Tree) classifiers over Naïve Bayes (NB) for spam filtering on YouTube(Alberto, Lochter, and Almeida 2015b). A comparative analysis of common YouTube comment spam filtering techniques show that high filtering accuracy (>= 98%) can be achieved with low-complexity algorithms (Abdullah et al. 2018). This study uses features based on the EdgeRank algorithm and is based on experiments employing nine different learning classifiers such as decision trees, Bayesian and function-based. The approach in (Chowdury et al. 2013)works with TubeKit API to crawl YouTube and create a dataset used to train their classifier to identify if the video is spam or not. It works with algorithms such as Naïve Bayes, Decision tree (D-Tree), and Clustering to label instances as spam or legitimate and compares the results from each approach to identify the most effective method of classification. (Alias, Foozy, and Ramli 2019) used six classifiers of machine learning techniques i.e Random Tree (RT), Random Forest (RF), Naive Bayes, KStar, Decision Tree, and Decision Stump for YouTube live streaming spam comments detection. Alberto, Lochter, and Almeida (Alberto, Lochter, and Almeida 2015a) introduces an online tool called TubeSpam for automatically detecting spam comments posted in the comments sections of YouTube videos. The authors also evaluate different state-of-the-art classification techniques and conclude that decision trees, logistic regression, Bernoulli naïve Bayes, random forests, and support vector machines are statistically equivalent. The most cited article was (O'Callaghan et al. 2012) in IEEE Explore which has 83 citations and 1556 full text views.

Our institution is keen on working on latest research trends and has extensive knowledge and research experience which resulted in quality publications (Rinesh et al. 2022; Sundararaman et al. 2022; Mohanavel et al. 2022; Ram et al. 2022; Dinesh Kumar et al. 2022; Vijayalakshmi et al. 2022; Sudhan et al. 2022; Kumar et al. 2022; Sathish et al. 2022; Mahesh et al. 2022; Yaashikaa et al. 2022). The main problem with this existing method is with a large data set, the prediction stage might be slow and accuracy depends on the quality of the data. In this paper, an automation technique for the prediction of spam comments in YouTube videos using the Decision Tree (D-Tree) method in comparison with the K-Nearest Neighbor (KNN) method. The Innovative Decision Tree algorithm is to better understand the effectiveness of spam comment prediction. The aim of this paper is to evaluate the accuracy of the Decision Tree (D-Tree) and K-Nearest Neighbor (K-NN) classifier before and after optimizing the important parameters using the Python software tool. The comparison of the two different models DecisionTree (D-Tree) and K-Nearest Neighbor (K-NN) are to be tested. The performance analysis of the proposed spam comment detection method gives better results than the existing K-Nearest Neighbor (K-NN) method.

## 2. Materials and Methods

This work was carried out in the Digital Image Processing Laboratory, Department of Computer

Science and Engineering, Saveetha School of Engineering. In this paper, The YouTube Spam Collection Data Set Collect from Kaggle(Lichman and Others 2013). The dataset contained ten selected videos and was downloaded from YouTube through API. It is composed of 1,956 real and non-encoded messages that were labeled as legitimate (ham) or spam. Each sample represents a text comment posted in the comments section of each selected video. Group 1 was a K-Nearest Neighbor (KNN) algorithm and Group 2 was an innovative Decision Tree (D-Tree) model. Python 3.9.2 and the NLP library were used for all implementations. The calculation is performed utilizing G-power 0.85 with alpha and beta qualities 0.05, 0.1 with a confidence interval at 95%. The steps involved in the implementation of the Innovative Decision Tree(D-Tree) algorithm are described as follows.

**K-Nearest Neighbor Algorithm**

The k-nearest-neighbors algorithm is a classification algorithm, and it is supervised: it takes a bunch of labeled points and uses them to learn how to label other points. To label a new point, it looks at the labeled points closest to that new point (those are its nearest neighbors), and has those neighbors vote, so whichever label the most of the neighbors have is the label for the new point (the "k" is the number of neighbors it checks). KNN algorithm is widely applied in pattern recognition and data mining for classification, which is famous for its simplicity and low error rate. A train data set with accurate classification labels should be known at the beginning of the algorithm. Then for a query data qi, whose label is not known and which is presented by a vector in the feature space, calculate the distances between it and every point in the train data set. After sorting the results of distance calculation, the decision of the class label of the test point qi can be made according to the label of the k nearest points in the train data set. The quality of the train data set directly affects the classification results. It is a nonlinear model which is built by many linear boundaries, here for a model that gives both labels and features so that it will understand to classify points based on features, due to overfitting in the data it is not accurate compared with other algorithms shown in Fig. 1.

The sample group 1 is the K-Nearest Neighbor (KNN) algorithm is a simple, supervised machine learning algorithm that can be used to solve both classification and regression problems. The steps involved in the implementation of the Innovative Decision Tree algorithm are described as follows.

**Decision Tree Algorithm**

The decision tree algorithm is based on Entropy, its main idea is to map all examples to different categories based upon different values of the condition attribute set; its core is to determine the best classification attribute from condition attribute sets. The algorithm chooses information gain as attribute selection criteria; usually the attribute that has the highest information gain is selected as the splitting attribute of the current node. Branches can be established based on different values of the attributes and the process above is recursively called on each branch to create other nodes and branches until all the samples in a branch belong to the same category. To select the splitting attributes, the concepts of Entropy and Information Gain and Gain Ratio are used. Entropy is a measure in information theory, which characterizes the impurity of an arbitrary collection of examples. If the target attribute takes on 'c' different values, then the entropy S relative to this c-wise classification is defined as

$$\text{Entropy}(s) = \Sigma - pi \log 2p \quad (1)$$

Where Pi is the probability of S belonging to class i. Logarithm is base 2 because entropy is a measure of the expected encoding length measured in bits. The decision tree is built in a top-down fashion. ID3 chooses the splitting attribute with the highest gain in information, where gain is defined as the difference between how much information is needed after the split. This is calculated by determining the differences between the entropies of the original data set and the weighted sum of the entropies from each of the subdivided data sets. The motive is to find the feature that best splits the target class into the purest possible children nodes - pure nodes with only one class. This measure of purity is called information. It represents the expected amount of information that would be needed to specify how a new instance of an attribute should be classified. The formula used for this purpose is:

$$G(D,S) = H(D) - \Sigma P(Di)H(Di) \quad (2)$$

The attribute with highest value of information gain is used as the splitting node thereby constructing the tree in top down fashion shown in Fig. 2.

The sample preparation group 2 is the novel Decision Tree (D-Tree) algorithm, which is considered to be one of the most useful Machine Learning algorithms since it can be used to solve a variety of problems. It can be used for classification and regression problems. The experimental results show that the proposed D-Tree method has achieved better accuracy results.

**Statistical Analysis**

The output is obtained by using Python software. To train these datasets, required a monitor with resolution of 1024×768 pixels (7th gen, i5, 4 8GB RAM, 500 GB HDD), and Python software with required library functions and tool functions. For statistical implementation, the software tool used here is IBM SPSS V26.0(Hilbe 2004). The independent sample t test was performed to find the mean, standard deviation and the standard error mean statistical significance between the groups, and then comparison of the two groups with the SPSS software will give the accurate values for the two different s which will be utilized with the graph to calculate the significant value with maximum accuracy value (94.47%), mean value (94%) and standard deviation value (0.84738). Dependent variables are accuracy and independent variables are KNN and D-Tree methods.

## 3. Results

Figure. 3 shows the simple bar graph for K-Nearest Neighbor (KNN) Classifier accuracy rate is compared with Decision Tree (D-Tree) Classifier. The Decision Tree (D-Tree) Classifier is higher in terms of accuracy rate 94.47% when compared with K-Nearest Neighbor (KNN) Classifier 86.91%. Variable results with its standard deviation ranging from 80 lower to 90 higher K-Nearest Neighbor (KNN) Classifier where Decision Tree (D-Tree) Classifier standard deviation ranging from 90 lower to 100 higher. There is a significant difference between K-Nearest Neighbor (KNN) Classifier and Decision Tree (D-Tree) Classifier ($p<0.05$ Independent sample test). X-axis: Decision Tree (D-Tree) Classifier accuracy rate vs K-Nearest Neighbor (KNN) Classifier Y-axis: Mean of accuracy rate, for identification of keywords ± 1 SD with 95 % CI.

Table.1 shows the Evaluation Metrics of Comparison of K-Nearest Neighbor (KNN) and Decision Tree (D-Tree) Classifier. The accuracy rate of K-Nearest Neighbor (KNN) is 86.91% and Decision Tree (D-Tree) is 94.47%. In all aspects of parameters Decision Tree (D-Tree) provides better performance compared with the K-Nearest Neighbor (KNN) of predicting Spam comments on YouTube videos with improved accuracy rate.

Table.2 shows the statistical calculation such as Mean, standard deviation and standard error Mean for K-Nearest Neighbor (KNN) and Decision Tree (D-Tree). The accuracy rate parameter used in the t-test. The mean accuracy rate of K-Nearest Neighbor (KNN) is 86.91% and Decision Tree (D-Tree) is 94.47%. The Standard Deviation of K-Nearest Neighbor (KNN) is 1.03829 and Decision Tree (D-Tree) is 0.84738. The Standard Error Mean of K-Nearest Neighbor (KNN) is 0.82939 and Decision Tree (D-Tree) is 0.18394.

Table.3 displays the statistical calculations for independent samples tested between K-Nearest Neighbor (KNN) and Decision Tree (D-Tree). The significance. The signal to noise ratio is 0.0291. Independent samples T-test is applied for comparison of K-Nearest Neighbor (KNN) and Decision Tree (D-Tree) with the confidence interval as 95% and level of significance as 0.33232. This independent sample test consists of significance as 0.001, significance (2-tailed), Mean difference, standard error difference, and lower and upper interval difference.

## 4. Discussion

A comparative study has been presented between decision tree and k-nearest neighbor models. Accuracy analysis has been performed to investigate the importance of each of the input parameters. D-Tree provides better accuracy output when compared to the KNN algorithm(Aziz, Foozy, and Shamala 2017). D-Tree is a powerful technique to classify human misbehavior activity. The accuracy result produced by D-Tree is better than the KNN method. D-Tree can significantly improve classification accuracy and time efficiency. This shows that the maximum accuracy is obtained quickly in the D-Tree algorithm. During the training process, the confusion matrix was used to evaluate the classification models (Thapa et al. 2021). The confusion matrix is a matrix that maps the predicted outputs across actual outputs. It is often used to describe the performance of a classification model on a set of test data. Important metrics were computed from the confusion matrix in order to evaluate the classification models(Ifriza and Sam'an 2021). In addition to correct classification rate or accuracy other metrics that were computed for evaluation were True Positive Rate (TPR), False Positive Rate (FPR), Precision, Accuracy, F1 score, and Misclassification rate.

Here I extracted the comments using the video URL and manually categorized them into four classes. Experiments were carried out to automatically categorize the extracted comments. The classified comments were evaluated using Precision (P), Recall (R) and Accuracy (A) metrics, estimated as below.

$$\text{Precision} = TP/(TP + FP) \quad (3)$$

$$\text{Recall} = TP/(TP + FN) \quad (4)$$

$$\text{Accuracy} = (TP + FN)/(TP + FP + TN + FN) \quad (5)$$

Precision is calculated as the ratio of the number of true positives to the total number of comments

classified to be true. Recall represents the ratio of the number of true positives to the total number of comments that are true and accuracy represents the total comments that are classified as true to the total number of comments that are classified (Han, Pei, and Kamber 2011). For relevant comments, TP indicates the number of relevant comments that are correctly classified as relevant, FP indicates the number of non-relevant comments that are incorrectly classified as relevant, TN represents the number of non-relevant comments that are correctly classified as non-relevant comments and FN represents the number of relevant comments that are incorrectly classified as non-relevant comments.

In classification, a total of two algorithms implemented in python were set as classifiers in detecting YouTube spam comments. The purpose of implementing the two algorithms is to compare accuracy. The classification of accuracy across two different classification algorithms such as Decision Tree (D-Tree) and k-Nearest Neighbor (KNN) using data proportion of 80:20 Ratio means 80% for training and 20% for testing. The result shows that the D-Tree classifier gives the highest accuracy when testing in python compared to the KNN classifier (Sadoon et al. 2017; Kanodia, Sasheendran, and Pathari 2018). The goal of this work is to find which algorithms provide high and best accuracy and precision to help in detecting Spam comments on YouTube (Sadoon et al. 2017). Using the D-Tree algorithm hope to overcome the disadvantages faced by the existing systems and achieve results that are more efficient and accurate in the classification of a given video as spam. The limitation of the proposed method is that the computational cost increases when the k value increases. Future work, with the deep neural network based implementations such as convolutional recurrent neural networks, may obtain better accuracy results for detecting unwanted YouTube Comments.

## 5. Conclusion

In this research, the development of a Youtube spam comment detection framework by using machine learning techniques has been done. It is important to improve security since the Internet nowadays indicates the security issues. The proposed model exhibits the K-Nearest Neighbor (KNN) and Decision Tree (D-Tree), in which the Decision Tree (D-Tree) has the highest values. The accuracy Rate of Decision Tree (D-Tree) is 94.47% is higher compared with K-Nearest Neighbor (KNN) that has an accuracy rate of 86.91% in analysis of detecting Spam comments on YouTube videos with improved accuracy rate.

**Declaration**

**Conflicts of Interest**

No conflict of interest in this manuscript

**Authors Contributions**

Author KR was involved in data collection, data analysis & manuscript writing. Author MVP was involved in conceptualization, data validation, and critical review of manuscripts.

**Acknowledgment**

The authors would like to express their gratitude towards Saveetha School of Engineering, Saveetha Institute of Medical And Technical Sciences (Formerly known as Saveetha University) for successfully carrying out this work.

**Funding:** We thank the following organizations for providing financial support that enabled us to complete the study.

1. Chipontime Technologies Pvt. Ltd. Bangalore.
2. Saveetha University
3. Saveetha Institute of Medical And Technical Sciences
4. Saveetha School of Engineering

## 6. References

Abdullah, Abdullah O., Mashhood A. Ali, Murat Karabatak, and Abdulkadir Sengur. 2018. "A Comparative Analysis of Common YouTube Comment Spam Filtering Techniques." In 2018 6th International Symposium on Digital Forensic and Security (ISDFS), 1–5. ieeexplore.ieee.org.

Aiyar, Shreyas, and Nisha P. Shetty. 2018. "N-Gram Assisted Youtube Spam Comment Detection." Procedia Computer Science 132 (January): 174–82.

Alberto, Túlio C., Johannes V. Lochter, and Tiago A. Almeida. 2015a. "Post or Block? Advances in Automatically Filtering Undesired Comments." Journal of Intelligent and Robotic Systems 80 (1): 245–59.

———. 2015b. "TubeSpam: Comment Spam Filtering on YouTube." In 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 138–43.

Alias, N., C. F. M. Foozy, and S. N. Ramli. 2019. "Video Spam Comment Features Selection Using Machine Learning Techniques." Indones. J. Electr. https://pdfs.semanticscholar.org/482c/9c66a70d72bed284d59740d6a26322d51b6c.pdf.

Alsaleh, Mansour, Abdulrahman Alarifi, Fatima

Al-Quayed, and Abdulmalik Al-Salman. 2015. "Combating Comment Spam with Machine Learning Approaches." In 2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA), 295–300.

Aziz, A., C. F. M. Foozy, and P. Shamala. 2017. "YouTube Spam Comment Detection Using Support Vector Machine and K–Nearest Neighbor." Indonesian Journal of Biotechnology. https://www.researchgate.net/profile/Cik-Feresa-Mohd-Foozy/publication/327249513_Youtube_spam_comment_detection_using_support_vector_machine_and_K-nearest_neighbor/links/5bcae524458515f7d9cbf793/Youtube-spam-comment-detection-using-support-vector-machine-and-K-nearest-neighbor.pdf.

Chowdury, Rashid, Md Nuruddin Monsur Adnan, G. A. N. Mahmud, and Rashedur M. Rahman. 2013. "A Data Mining Based Spam Detection System for YouTube." In Eighth International Conference on Digital Information Management (ICDIM 2013), 373–78.

Dinesh Kumar, M., V. Godvin Sharmila, Gopalakrishnan Kumar, Jeong-Hoon Park, Siham Yousuf Al-Qaradawi, and J. Rajesh Banu. 2022. "Surfactant Induced Microwave Disintegration for Enhanced Biohydrogen Production from Macroalgae Biomass: Thermodynamics and Energetics." Bioresource Technology 350 (April): 126904.

Hamou, Reda Mohamed, and Abdelmalek Amine. 2013. "The Impact of the Mode of Data Representation for the Result Quality of the Detection and Filtering of Spam." International Journal of Information Retrieval Research (IJIRR) 3 (1): 43–59.

Han, Jiawei, Jian Pei, and Micheline Kamber. 2011. Data Mining: Concepts and Techniques. Elsevier.

Hilbe, Joseph M. 2004. "A Review of SPSS 12.01, Part 2." The American Statistician 58 (2): 168–71.

Ifriza, Yahya Nur, and Muhammad Sam'an. 2021. "Performance Comparison of Support Vector Machine and Gaussian Naive Bayes Classifier for Youtube Spam Comment Detection." Journal of Soft Computing Exploration 2 (2): 93–98.

Jin, Xin, Cindy Xide Lin, Jiebo Luo, and Jiawei Han. 2011. "SocialSpamGuard: A Data Mining-Based Spam Detection System for Social Media Networks." Proceedings of the VLDB Endowment International Conference on Very Large Data Bases 4 (12): 1458–61.

Kanodia, Simran, Rachna Sasheendran, and Vinod Pathari. 2018. "A Novel Approach for Youtube Video Spam Detection Using Markov Decision Process." 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI). https://doi.org/10.1109/icacci.2018.8554405.

Kumar, J. Aravind, J. Aravind Kumar, S. Sathish, T. Krithiga, T. R. Praveenkumar, S. Lokesh, D. Prabu, A. Annam Renita, P. Prakash, and M. Rajasimman. 2022. "A Comprehensive Review on Bio-Hydrogen Production from Brewery Industrial Wastewater and Its Treatment Methodologies." Fuel. https://doi.org/10.1016/j.fuel.2022.123594.

Lichman, Moshe, and Others. 2013. "UCI Machine Learning Repository, 2013." URL Http://archive. Ics. Uci. Edu/ml 40.

Mahesh, Narayanan, Srinivasan Balakumar, Uthaman Danya, Shanmugasundaram Shyamalagowri, Palanisamy Suresh Babu, Jeyaseelan Aravind, Murugesan Kamaraj, and Muthusamy Govarthanan. 2022. "A Review on Mitigation of Emerging Contaminants in an Aqueous Environment Using Microbial Bio-Machines as Sustainable Tools: Progress and Limitations." Journal of Water Process Engineering. https://doi.org/10.1016/j.jwpe.2022.102712.

Mohanavel, Vinayagam, K. Ravi Kumar, T. Sathish, Palanivel Velmurugan, Alagar Karthick, M. Ravichandran, Saleh Alfarraj, Hesham S. Almoallim, Shanmugam Sureshkumar, and J. Isaac JoshuaRamesh Lalvani. 2022. "Investigation on Inorganic Salts K2TiF6 and KBF4 to Develop Nanoparticles Based TiB2 Reinforcement Aluminium Composites." Bioinorganic Chemistry and Applications 2022 (January): 8559402.

O'Callaghan, Derek, Martin Harrigan, Joe Carthy, and Pádraig Cunningham. 2012. "Network Analysis of Recurring Youtube Spam Campaigns." In Sixth International AAAI Conference on Weblogs and Social Media. aaai.org. https://www.aaai.org/ocs/index.php/ICWSM/ICWSM12/paper/viewPaper/4579.

Ram, G. Dinesh, G. Dinesh Ram, S. Praveen Kumar, T. Yuvaraj, Thanikanti Sudhakar Babu, and Karthik Balasubramanian. 2022. "Simulation and Investigation of MEMS Bilayer Solar Energy Harvester for Smart Wireless Sensor Applications." Sustainable Energy Technologies and Assessments.

https://doi.org/10.1016/j.seta.2022.102102.

Rinesh, S., K. Maheswari, B. Arthi, P. Sherubha, A. Vijay, S. Sridhar, T. Rajendran, and Yosef Asrat Waji. 2022. "Investigations on Brain Tumor Classification Using Hybrid Machine Learning Algorithms." Journal of Healthcare Engineering 2022 (February): 2761847.

Sadoon, Fahad M. Al, Fahad M. Al Sadoon, Muhammad Qasim, and Naif A. Darwish. 2017. "PREDICTION OF THERMODYNAMIC STABILITY LIMITS AND CRITICALITY CONDITIONS FOR BINARY HYDROCARBON SYSTEMS." Computational Thermal Sciences: An International Journal. https://doi.org/10.1615/computthermalscien.2017018774.

Sathish, T., V. Mohanavel, M. Arunkumar, K. Rajan, Manzoore Elahi M. Soudagar, M. A. Mujtaba, Saleh H. Salmen, Sami Al Obaid, H. Fayaz, and S. Sivakumar. 2022. "Utilization of Azadirachta Indica Biodiesel, Ethanol and Diesel Blends for Diesel Engine Applications with Engine Emission Profile." Fuel. https://doi.org/10.1016/j.fuel.2022.123798.

Schultes, Peter, Verena Dorner, and Franz Lehner. 2013. "Leave a Comment! An In-Depth Analysis of User Comments on YouTube." Wirtschaftsinformatik 42: 659–73.

Sudhan, M. B., M. Sinthuja, S. Pravinth Raja, J. Amutharaj, G. Charlyn Pushpa Latha, S. Sheeba Rachel, T. Anitha, T. Rajendran, and Yosef Asrat Waji. 2022. "Segmentation and Classification of Glaucoma Using U-Net with Deep Learning Model." Journal of Healthcare Engineering 2022 (February): 1601354.

Sundararaman, Sathish, J. Aravind Kumar, Prabu Deivasigamani, and Yuvarajan Devarajan. 2022. "Emerging Pharma Residue Contaminants: Occurrence, Monitoring, Risk and Fate Assessment – A Challenge to Water Resource Management." Science of The Total Environment. https://doi.org/10.1016/j.scitotenv.2022.153897.

Thapa, Rahul, Bikal Lamichhane, Dongning Ma, and Xun Jiao. 2021. "SpamHD: Memory-Efficient Text Spam Detection Using Brain-Inspired Hyperdimensional Computing." In 2021 IEEE Computer Society Annual Symposium on VLSI (ISVLSI), 84–89. ieeexplore.ieee.org.

Vijayalakshmi, V. J., Prakash Arumugam, A. Ananthi Christy, and R. Brindha. 2022. "Simultaneous Allocation of EV Charging Stations and Renewable Energy Sources: An Elite RERNN-m2MPA Approach." International Journal of Energy Research. https://doi.org/10.1002/er.7780.

Yaashikaa, P. R., P. Senthil Kumar, S. Jeevanantham, and R. Saravanan. 2022. "A Review on Bioremediation Approach for Heavy Metal Detoxification and Accumulation in Plants." Environmental Pollution 301 (May): 119035.

Narayanasamy, S., Sundaram, V., Sundaram, T., & Vo, D. V. N. (2022). Biosorptive ascendency of plant based biosorbents in removing hexavalent chromium from aqueous solutions–Insights into isotherm and kinetic studies. Environmental Research, 210, 112902.

**Tables And Figures**

Table.1. Comparison of K-Nearest Neighbor (KNN) and Decision Tree (D-Tree) Classifier for predicting the spam comments in YouTube videos with improved accuracy rate. The accuracy rate of K-Nearest Neighbor (KNN) is 86.91 and Decision Tree (D-Tree) has 94.47.

| SI.No. | Test Size | Accuracy Rate | |
|---|---|---|---|
| | | K-Nearest Neighbor | Decision Tree |
| 1 | Test 1 | 84.23 | 90.10 |
| 2 | Test2 | 84.43 | 90.50 |
| 3 | Test3 | 84.87 | 90.77 |
| 4 | Test4 | 85.12 | 91.92 |

| | | | |
|---|---|---|---|
| 5 | Test5 | 85.23 | 92.42 |
| 6 | Test6 | 85.34 | 92.91 |
| 7 | Test7 | 86.27 | 93.65 |
| 8 | Test8 | 86.42 | 93.88 |
| 9 | Test9 | 86.78 | 94.18 |
| 10 | Test10 | 86.90 | 94.34 |

Table .2. The statistical calculation such as Mean, standard deviation and standard error Mean for K-Nearest Neighbor (KNN) and Decision Tree (D-Tree). The accuracy rate parameter used in the t-test. The mean accuracy rate of K-Nearest Neighbor (KNN) is 86.91 and Decision Tree (D-Tree) is 94.47. The Standard Deviation of K-Nearest Neighbor (KNN) is 1.03829 and the Decision Tree (D-Tree) is 0.84738. The Standard Error Mean of K-Nearest Neighbor (KNN) is 0.82939 and the Decision Tree (D-Tree) is 0.18394.

| **Group** | | **N** | **Median** | **Standard Deviation** | **Standard Error Median** |
|---|---|---|---|---|---|
| Accuracy | Decision Tree | 10 | 94.47 | 0.84738 | 0.18394 |
| | K-Nearest Neighbor (Knn) | 10 | 86.91 | 1.03829 | 0.82939 |

Table. 3. The statistical calculations for independent samples tested between K-Nearest Neighbor (KNN) and Decision Tree (D-Tree). The significance for signal to noise ratio is 0.0291 Independent samples T-test is applied for comparison of K-Nearest Neighbor (KNN) and Decision Tree(D-Tree) with the confidence interval as 95% and level of Significance as 0.33232, This independent sample test consists of significance as 0.001, significance (2-tailed), Mean difference, standard error difference, and lower and upper interval difference.

| **Group** | **Levene,s Test for Equality of Variances** | | **T-Test for Equality of Means** | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | **F** | **sig.** | **t** | **df** | **sig(2-tailed)** | **Mean Difference** | **Std Error Difference** | **95% Confidence Interval (Lower)** | **95% Confidence Interva1 (Upper)** |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Accuracy | Equal Variances assumed | 8.384 | 0.0291 | 16.237 | 18 | .001 | 13.40304 | 0.98273 | 12.91833 | 15.2343 |
| | Equal Variances assumed | | | 12.432 | 11.129 | .001 | 12.18293 | 0.12834 | 11.23422 | 14.0344 |

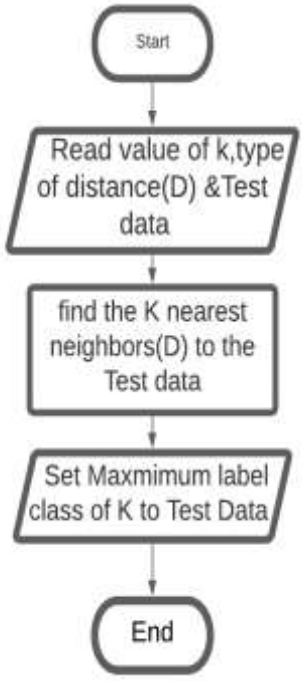


Fig.1. Flow Chart of K-Nearest Neighbor (KNN)

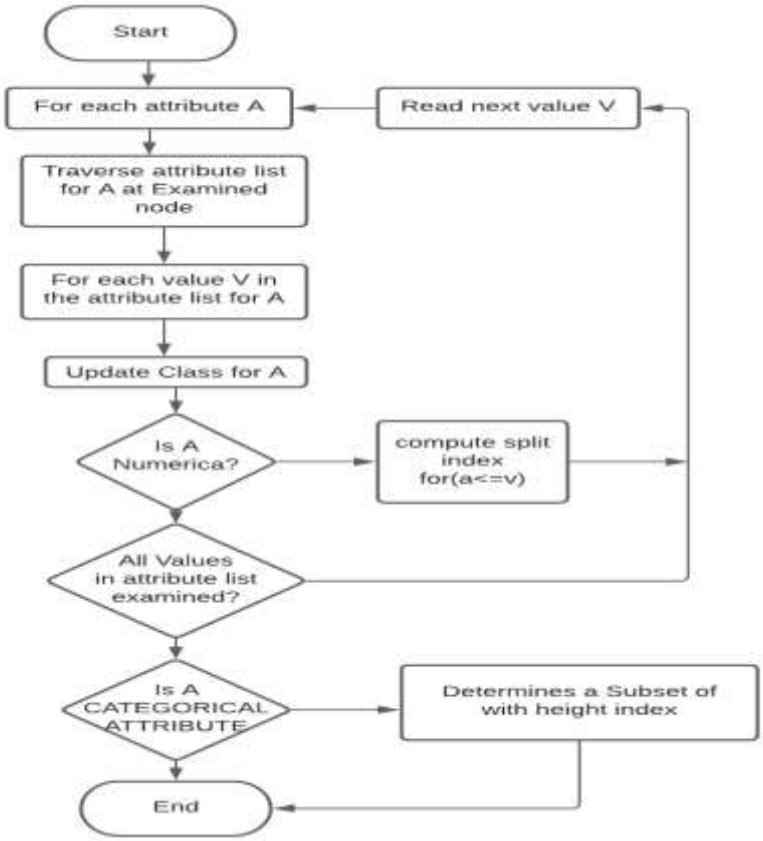


Fig. 2 Flow Chart of Decision Tree (D-Tree)

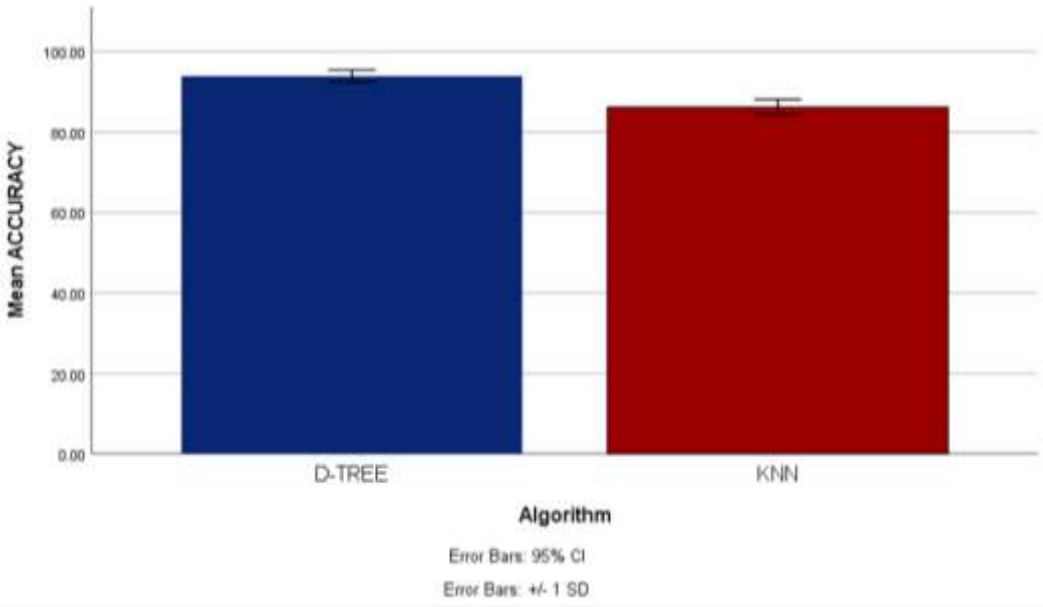


Fig. 3. Bar graph between KNN and Innovative Decision Tree Classifier. Comparison of KNN algorithm and DT in terms of mean accuracy. The mean accuracy of KNN is better than NB and the standard deviation of KNN is slightly better than DT. X-Axis: KNN vs DT Y-Axis: Mean accuracy of detection ± 1 SD with CI of 95%.