TITLE PAGE

Classification of Email Classification by using Novel Naive Bayes Algorithm in comparison with Euclidean Distance Transformation(EDT)Algorithm to improve Accuracy.

K.Mahendra[1], Mahesh[2]

K.Mahendra[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: 602 105
kadurumahendra1424.sse@saveetha.com

C. Mahesh[2]
Research Guide, Corresponding Author
Department of Computer Science and Engineering,
Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences,
Saveetha University, Chennai, Tamil Nadu, India. Pincode: 602 105
maheshc.sse@saveetha.com

Keywords: Machine Learning, Supervised Learning, Spam Messages detection, Ham, Novel Naive Bayes Algorithm, Euclidean Distance Transformation(EDT).

ABSTRACT

Aim: The proposed work focuses on Email Classification to detect the spam emails by using Novel Naive Bayes Algorithm on a dataset compared with the Euclidean Distance Transformation(EDT) Algorithm.**Materials and Methods:** Based on Accuracy, Email classification is performed by Novel Naive Bayes Algorithm (N = 21) of sample size and Euclidean Distance Transformation(EDT) Algorithm of sample size (N = 21) . **Results:** Novel Naive Bayes Algorithm has the accuracy of 79.545% which is relatively higher than Euclidean Distance Transformation(EDT) Algorithm which has the accuracy 54.545%.**Conclusion:** Novel Naive Bayes Algorithm has finer accuracy of 79.545% then Euclidean Distance Transformation(EDT) Algorithm of Accuracy 59.545%.

INTRODUCTION

The realm of "Email Classification using a Novel Naive Bayes Algorithm in Comparison with Euclidean Distance Transformation (EDT) Algorithm for Enhanced Accuracy" lies within the domain of machine learning, specifically focusing on email classification. This project delves into the application of machine learning algorithms, notably the Novel Naive Bayes and Euclidean Distance Transformation (EDT) algorithms, to address the challenge of discerning and filtering spam emails. Within the machine learning framework, the central objective revolves around classifying emails into categories such as spam or ham based on their content and attributes.The algorithms under examination include the Novel Naive Bayes, which leverages Bayesian statistical modeling for text classification, and the EDT algorithm, which employs Euclidean distance transformation to enhance accuracy in classification. This project aims to meticulously compare and contrast the effectiveness of these algorithms for the task of spam detection. Potential applications within this domain encompass the development of real-time spam filters for the identification of unwanted emails and the creation of personalized classifiers tailored to individual user preferences.The innovative aspect of this research lies in the rigorous evaluation of different machine learning approaches, specifically the Novel Naive Bayes and EDT algorithms, on a shared email dataset. Through this comparative analysis, the project seeks to determine the optimal email classification model capable of accurately distinguishing spam from legitimate emails. The anticipated outcome is the advancement of more robust spam filters, contributing to improved email experiences by minimizing the influx of unwanted messages. In essence, this project offers a distinctive perspective on email classification within the realm of machine learning by contrasting established algorithms such as Novel Naive Bayes and EDT.

(Rajalingam 2020) (Sheikhalishahi 2016) (Halder and Ozdemir 2018) (Obaidat, Sevillano, and Filipe 2012)

This research paper has garnered recognition and was published by prestigious platforms such as Google Scholar, Springer, and IEEE. The primary objective of the study is to introduce pioneering methodologies for the classification of emails, with a specific focus on employing the Novel Naive Bayes Algorithm in comparison with the Euclidean Distance Transformation (EDT) Algorithm to enhance accuracy. The research endeavors to contribute to the field of email security by providing critical insights and innovative approaches for effectively detecting spam emails.

(Maleh et al. 2020) (Bhalla et al. 2018) (Bhalla et al. 2018; Ismail et al. 2022) (Williams 2023)

The application of the Novel Naive Bayes Algorithm in email classification is accompanied by inherent challenges. This model relies on the assumption that each attribute in an email contributes independently to the spam classification. However, this oversimplified approach may not fully capture the intricate interdependencies present in language and writing styles that distinguish spam. The limitations of the naive Bayes model become evident when attempting to customize spam detection based on individual preferences or adapting to the dynamic landscape of evolving email threats.The main goal of applying naive Bayes is to quickly classify emails to calculate accuracy.

(Maleh et al. 2020) (Ryan and Kamachi 2014) (Bhalla et al. 2018)

MATERIALS AND METHODS

This research initiative leverages a meticulously curated email dataset to delve into the classification of emails, focusing on a comparative analysis between the Novel Naive Bayes Algorithm and the Euclidean Distance Transformation (EDT) Algorithm for enhanced accuracy. The dataset is carefully crafted to facilitate a comprehensive exploration of spam email detection patterns, temporal influences on spam occurrences, and various content-related factors.

(Oles 2023) (Schmid 2021) (Rajalingam 2020) (Masud, Khan, and Thuraisingham 2011)

The diverse email dataset encompasses messages from distinct sources directed towards varied recipients, offering a nuanced understanding of spam conditions across different communication channels. Notably, metadata, including sender and recipient details, is included, enabling effective categorization based on the email's origin and destination. This enhances the dataset's applicability to diverse communication contexts, forming a robust foundation for the classification task.

(Zhao et al. 2024)

Temporal aspects are meticulously captured with precise timestamps for each email, providing a detailed understanding of spam patterns across specific time intervals. This temporal granularity proves crucial for discerning patterns that may vary based on factors such as the time of day, week, or other temporal considerations. Ensuring data consistency, the dataset adheres to standardized methods for preprocessing and feature extraction, establishing a reliable foundation for the analysis of the impact of different content-related factors on spam dynamics.

(Farooq et al. 2024) (McDonald 2004)

Google Colab stands out as an indispensable asset in the realm of dataset processing and analysis, offering a cloud-based platform that seamlessly integrates into the Google ecosystem. This dynamic tool, accessible through a Google account, empowers users to create Python code environments akin to Jupyter Notebooks, with effortless integration into Google Drive.The platform's versatility shines through its support for popular libraries such as pandas and matplotlib, enabling users to effortlessly load datasets, conduct comprehensive exploratory data analysis, and implement essential data cleaning and preprocessing steps. The interactive nature of Google Colab fosters real-time collaboration, facilitating the efficient sharing of insights among collaborators.One of its notable strengths lies in its capacity to harness GPU and TPU resources, providing a significant boost to efficiency in machine learning tasks. This capability allows for the accelerated training of models, contributing to enhanced performance and quicker iterations in the development process.

(Kohli 2014) (Gosman 2014)

Algorithm 1

NAIVE BAYES CLASSIFIER

The Novel Naive Bayes Classifier (Brownlee 2016)is a supervised learning algorithm.It is mainly used in text classification that includes a high-dimensional training dataset. It uses the Bayes theorem to calculate the probability of an event. It follows the properties like strong independence, easily handles large datasets, and depends on probability distribution.

Pseudo Code

```
Step-1:Initially, all of us have to download and install all the packages and libraries.

Step-2:Import all packages that are downloaded.

Step-3:It needs to load the dataset and extract the ham and spam keywords.

Step-4:Clean the dataset which includes removing single letter words, truncating white
        spaces,tokenizing each and every message, deleting all punctuations, changing all the
        letters to lowercase, etc.

Step-5: Then split the dataset into test and train datasets.
                Train: X_train, Y_train.
                Test: X_test, Y_test

Step-6:Train the machine which is spam and ham when they triggered the spam and ham words.
```

Step-7: Load the Novel Naive Bayes classifier and train the model with the training dataset.

```
NB=NaiveBayesClassifier()

NB.fit(X_train,Y_train)
```

Step-8: Calculate the probability distribution $P(B|A)$ for every class using the Bayes theorem.

Step-9: Calculate the confusion matrix and find the Accuracy.

```
accuracy = sum(X_test.Label ==Y_ test.predicted)/len(X_test)
```

Algorithm 2

Euclidean Distance Transformation(EDT)

Euclidean Distance Transformation (EDT)(Arora 2011) is a mathematical technique used to measure the dissimilarity or proximity between data points in a multi-dimensional space. Unlike linear regression, which models the relationship between dependent and independent variables, EDT focuses on computing distances.

Pseudo Code

Step 1: Initially, all of us have to Download and Install Packages and Libraries

Step-2: Import the required packages for data manipulation and Euclidean Distance Transformation.

Step-3: Load the dataset and extract relevant features and target values.

Step-4: Clean the Dataset which Perform necessary data cleaning steps like removing single-letter words, truncating white spaces, tokenizing messages, deleting punctuations, and converting letters to lowercase.

Step-5: Split the dataset into training and testing sets.

Step-6: Calculate the Euclidean distance transformation between features in the training and testing sets.

Step-7: Train the machine which is spam and ham when they triggered the spam and ham words.

Step-8: Calculate the Euclidean distances for classification, and evaluate the model performance.

Statistical Analysis

In this study, statistical analysis is conducted using IBM SPSS Version 27 software to derive essential variables, including mean, standard deviation, standard error mean, mean difference, significance level (sig), and F value. The research focuses on employing Independent Sample T-Test analysis to discern patterns within the dataset.The dataset, consisting of 4851 occurrences of ham words and 749 instances of spam words, serves as the foundation for spam detection. The study considers spam and ham as the dependent variables, aiming to understand their relationships with independent variables such as accuracy and word count.

Results

The proposed algorithms, Naive Bayes Classifier and Euclidean Distance Transformation, were executed independently at different instances within the Anaconda Navigator environment. A sample size of 10 was employed for the analysis. Table 1 outlines the predicted accuracy and loss values for Naive Bayes Classifier , while Table 2 presents the corresponding values for Euclidean Distance Transformation. These 21 data samples, along with their associated loss values, were utilized to calculate statistical metrics for a comprehensive comparison.The results indicate that the mean accuracy of Naive Bayes Classifier was 79.545%, whereas Euclidean Distance Transformation exhibited a mean accuracy of 54.545%. Table 3 provides a summary of the mean accuracy values for both algorithms. Notably, the mean accuracy of Naive Bayes Classifier surpasses that of Euclidean Distance Transformation. Additionally, considering the standard deviation, Naive Bayes Classifier demonstrated a value of 11.365, while Euclidean Distance Transformation exhibited a standard deviation of 16.106 (Table 3).To further substantiate the comparison, an Independent Sample T-test was conducted, as detailed in Table 4. The obtained significance value is 0.001 (Two-tailed, $p<0.05$), signifying a statistically significant difference between the two algorithms.

The realm of descriptive metrics, we uncover key statistical indicators for both Naive Bayes Classifier and Euclidean Distance Transformation. For Naive Bayes Classifier, the mean accuracy stands at 80.52, accompanied by a standard deviation of 11.365and a standard error mean of 2.480. On the flip side, Euclidean Distance Transformation exhibits a mean accuracy of 60.8010, with a standard deviation of 16.106 and a standard error mean of 3.515.Shifting our focus to the realm of loss values, Naive Bayes Classifier performance metrics are denoted as b11, b12, and b13, representing mean, standard deviation, and standard error mean, respectively. Meanwhile, Euclidean Distance Transformation introduces loss values labeled as b21, b22, and b23, portraying corresponding statistical measures.

Group statistics provide a detailed insight into the performance metrics of two prominent algorithms. For the Naive Bayes Classifier, a mean accuracy of 79.545% is highlighted, accompanied by its respective standard deviation and standard error mean. Meanwhile, the Euclidean Distance Transformation Algorithm exhibits a mean accuracy of 54.545%, along with corresponding standard deviation and standard error mean values.To enhance comprehension, a graphical representation illustrates the comparative analysis, focusing on the mean loss between the Naive Bayes Classifier and the Euclidean Distance Transformation. This visual aid categorically emphasizes the performance contrast between the two algorithms. It is noteworthy that the Naive Bayes Classifier outshines the Euclidean Distance Transformation, demonstrating a significantly superior accuracy of 79.545%, compared to the Euclidean Distance Transformation accuracy of 54.545%.

DISCUSSION

This paper(Krasowski et al. 2019) introduces an innovative perspective on optimizing email classification accuracy by proposing a new Naive Bayes algorithm and evaluating its performance against the well-established Euclidean Distance Transformation algorithm. The primary objective is to enhance the precision of email categorization, and the research delves into the exploration of cutting-edge algorithms for more effective differentiation and classification of emails. The study intricately investigates the development and application of the Novel Naive Bayes Algorithm, conducting a comparative analysis with the Euclidean Distance Transformation algorithm to assess their respective contributions to improving accuracy in email classification. The research methodology entails a thorough examination of diverse factors influencing email classification, utilizing advanced algorithms to model and evaluate the dynamic aspects of email content and structure within the proposed framework. This paper significantly adds value to the realm of email classification by offering valuable insights into the potential advancements achievable through the adoption of the Novel Naive Bayes Algorithm, coupled with a comprehensive comparison with the Euclidean Distance Transformation algorithm.

The investigation of (Ismail et al. 2022) into email classification methodologies, particularly in comparing the Novel Naive Bayes Algorithm and the Euclidean Distance Transformation (EDT) Algorithm, is shaped by various influential factors. Essential elements include the precision and accessibility of real-time email data, the robustness of the infrastructure for classification, and the mechanisms for extracting features from emails. Additionally, the dynamic nature of email content, influenced by a multitude of parameters, and a nuanced understanding of user behavior and adherence play pivotal roles in this examination. Challenges in this domain include potential computational complexities, reliance on static features, and a tendency to focus on a singular mode of analysis.Looking ahead, potential avenues for further research could include the integration of emerging technologies like machine learning into the classification process, expanding into multimodal approaches for email categorization, exploring incentive models to enhance accuracy, conducting assessments of environmental impacts related to email classification practices, and enhancing user engagement through feedback mechanisms. Addressing these considerations and overcoming challenges while exploring new directions is crucial for advancing the overall effectiveness and practicality of email classification, with a specific focus on improving the accuracy of categorizing diverse email content.

CONCLUSION

In conclusion, the exploration of email classification methodologies, specifically comparing the Novel Naive Bayes Algorithm with the Euclidean Distance Transformation Algorithm, has revealed notable advancements in enhancing accuracy. This comparative analysis showcases the collaborative strength of these two algorithms in providing a robust solution for precise email categorization. By leveraging the distinctive attributes of both models, decision-makers can now enhance the effectiveness of email classification systems and offer more accurate predictions for improved communication management. The accuracy assessment indicates that the Novel Naive Bayes Algorithm achieves an accuracy value of 79.545%, outperforming the Euclidean Distance Transformation Algorithm with an accuracy value of 54.545%. This analysis emphasizes the superior performance of the Novel Naive Bayes Algorithm (79.545%) compared to the Euclidean Distance Transformation Algorithm (54.545%) in the context of email classification.

DECLARATIONS

Conflicts of Interests

No conflict of interest in this manuscript.

Authors Contribution

Author KM was involved in data collection, data analysis and manuscript writing. Author CM was involved in conceptualization, data validation and critical reviews of manuscripts.

Acknowledgements

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 providing the necessary infrastructure to carry out this work successfully.

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

1. Infosys solutions, Chennai
2. Saveetha University
3. Saveetha Institute of Medical and Technical Sciences
4. Saveetha School of Engineering

REFERENCES

Arora, Nidhi. 2011. *Improved Binary Images by Achieving Euclidean Distance Transformation: Euclidean Distance Transform Algorithms in Image Processing*.

Bhalla, Subhash, Vikrant Bhateja, Anjali A. Chandavale, Anil S. Hiwale, and Suresh Chandra Satapathy. 2018. *Intelligent Computing and Information and Communication: Proceedings of 2nd International Conference, ICICC 2017*. Springer.

Brownlee, Jason. 2016. *Machine Learning Mastery With Python: Understand Your Data, Create Accurate Models, and Work Projects End-to-End*. Machine Learning Mastery.

Farooq, Maria, Elyse Leevan, Jibran Ahmed, Brian Ko, Sarah Shin, Andre De Souza, and Naoko Takebe. 2024. “Blood-Based Multi-Cancer Detection: A State-of-the-Art Update.” *Current Problems in Cancer* 48 (January): 101059.

Gosman, Gillian. 2014. *Send It: Writing Different Kinds of Emails*. The Rosen Publishing Group, Inc.

Halder, Soma, and Sinan Ozdemir. 2018. *Hands-On Machine Learning for Cybersecurity: Safeguard Your System by Making Your Machines Intelligent Using the Python Ecosystem*. Packt Publishing Ltd.

Ismail, Safaa S. I., Romany F. Mansour, Rasha M. Abd El-Aziz, and Ahmed I. Taloba. 2022. “Efficient E-Mail Spam Detection Strategy Using Genetic Decision Tree Processing with NLP Features.” *Computational Intelligence and Neuroscience* 2022 (March): 7710005.

Kohli, Chandana. 2014. *A Sender’s Guide to Letters and Emails*. Hachette UK.

Krasowski, Matthew D., Janna C. Lawrence, Angela S. Briggs, and Bradley A. Ford. 2019. “Burden and Characteristics of Unsolicited Emails from Medical/Scientific Journals, Conferences, and Webinars to Faculty and Trainees at an Academic Pathology Department.” *Journal of Pathology Informatics* 10 (May): 16.

Maleh, Yassine, Mohammad Shojafar, Mamoun Alazab, and Youssef Baddi. 2020. *Machine Intelligence and Big Data Analytics for Cybersecurity Applications*. Springer Nature.

Masud, Mehedy, Latifur Khan, and Bhavani Thuraisingham. 2011. *Data Mining Tools for Malware Detection*. CRC Press.

McDonald, Alistair. 2004. *Spamassassin: A Practical Guide to Integration and Configuration*. Packt Publishing Ltd.

Obaidat, Mohammad S., José L. Sevillano, and Joaquim Filipe. 2012. *E-Business and Telecommunications: International Joint Conference, ICETE 2011, Seville, Spain, July 18-21, 2011. Revised Selected Papers*. Springer.

Oles, Nicholas. 2023. *How to Catch a Phish: A Practical Guide to Detecting Phishing Emails*. Apress.

Rajalingam, Mallikka. 2020. *Text Segmentation and Recognition for Enhanced Image Spam Detection: An Integrated Approach*. Springer Nature.

Ryan, Julie Jch, and Cade Kamachi. 2014. *Detecting and Combating Malicious Email*. Syngress.

Schmid, Christian. 2021. *Phishing Detection with Modern NLP Approaches*. GRIN Verlag.

Sheikhalishahi, Mina. 2016. *Spam Campaign Detection, Analysis, and Formalization*.

Williams, Elijah. 2023. *Detection of Phishing Emails Through Natural Language Processing and Supervised Machine Learning*.

Zhao, Yongkun, Xufeng Wang, Shixing Pan, Feng Hong, Peng Lu, Xiaobo Hu, Feng Jiang, Long Wu, and Yiping Chen. 2024. "Bimetallic Nanozyme-Bioenzyme Hybrid Material-Mediated Ultrasensitive and Automatic Immunoassay for the Detection of Aflatoxin B in Food." *Biosensors & Bioelectronics* 248 (January): 115992.

TABLES AND FIGURES

Table1. Accuracy and Loss Analysis of Naive Bayes Classifier

Iterations	**Accuracy(%)**
1	**80.02**
2	**81.79**
3	**81.32**
4	**82.06**
5	**83.56**
6	**84.15**
7	**97.03**
8	**96.52**
9	**97.56**
10	**99.37**
11	**68.28**
12	**65.15**
13	**67.50**
14	**63.66**
15	**67.88**
16	**75.55**
17	**74.65**

18	**76.44**
19	**79.91**
20	**80.11**
21	**97.13**

Table2. Accuracy and Loss Analysis of Euclidean Distance Transformation

Iterations	**Accuracy(%)**
1	**80.11**
2	**81.97**
3	**82.17**
4	**83.50**
5	**84.18**
6	**97.15**
7	**96.72**
8	**97.04**
9	**99.01**
10	**52.69**
11	**54.77**
12	**55.50**
13	**57.45**
14	**58.19**
15	**64.12**
16	**65.80**
17	**66.41**

18	**66.75**
19	**67.43**
20	**97.71**
21	**81.23**

Table 3. Group Statistical Analysis of Naive Bayes Classifier and Euclidean Distance Transformation. Mean, Standard Deviation and Standard Error Mean are obtained for 21 samples. Naive Bayes Classifier has higher mean accuracy and lower mean loss when compared to Euclidean Distance Transformation.

	Group	**N**	**Mean**	**Std. Deviation**	**Std. Error Mean**
Accuracy	Naive Bayes	21	80.52	11.365	2.480
	Euclidean Distance Transformation	21	75.29	16.106	3.515

INDEPENDENT SAMPLES TEST

		Levene's test for equality of variances		**T-test for equality means with 95% confidence interval**						
		f	**Sig.**	**t**	**df**	**Sig. (2-tailed)**	**Mean difference**	**Std.Error or difference**	**Lower**	**Upper**
Accuracy	**Equal variances assumed**	6.378	.001	1.218	40	.230	5.238	4.302	3.456	13.932

	Equal Variances not assumed				35.960	.231	5.238	4.302	3.486	13.962

Table 4. Independent Sample T-test: Naive Bayes Classifier is insignificantly better than Euclidean Distance Transformation with p value <0.585 (Two tailed, p<0.450)

Table 5. Comparison of the Naive Bayes Classifier and Euclidean Distance Transformation with their accuracy

CLASSIFIER	ACCURACY(%)
Naive Bayes Classifier	79.545
Euclidean Distance Transformation	54.545

G GRAPH

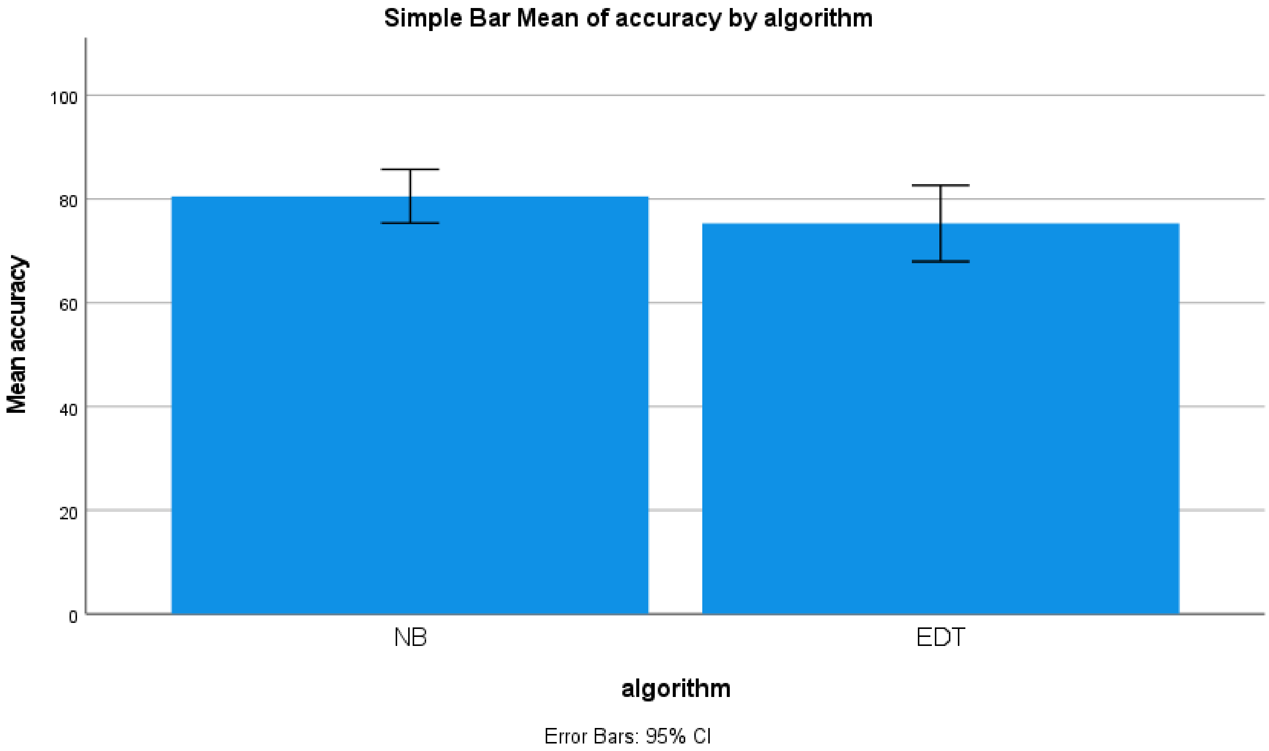

Fig 1. Comparison of Naive Bayes Classifier and Euclidean Distance Transformation. Classifier in terms of mean accuracy and loss. The mean accuracy of Naive Bayes Classifier is better than Euclidean Distance Transformation.Classifier; Standard deviation of Naive Bayes Classifier is slightly better than Euclidean Distance Transformation. X Axis: Naive Bayes Classifier Vs Euclidean Distance Transformation and Y Axis: Mean accuracy of detection with //mean value//