Smoke Detection

Name	ID	CRN
Mostafa Tarek	94071	6314
Mohamed Ezz	94303	6314
Abdelrahman Ahmed	94218	6316

Doctor: Muhammad Deif
ENG: Shaimaa Bahaa

Abstract

This research presents a comprehensive study on the application of diverse machine learning models for the task of Smoke Detection. The study encompasses a wide range of classifiers, including Support Vector Machines (SVM), XGBoost (XGB), Logistic Regression, k-nearest Neighbors (KNN), Neural Networks (NN), Decision Trees, and Random Forests. Each model is evaluated on its ability to accurately classify instances of smoke, contributing to the understanding of their strengths and weaknesses in the context of smoke detection.

Furthermore, the investigation extends to the impact of different feature selection techniques on model performance. Various strategies, such as variance thresholding, correlation analysis, feed-forward selection, and k-best selection, are employed to identify and retain the most relevant features for classification. The goal is to discern the influence of feature selection on the models' predictive capabilities and computational efficiency.

The experimentation involves rigorous testing and comparison of the models on benchmark datasets, with a focus on metrics such as accuracy, precision, recall, and F1 score. Insights gained from these comparisons are

crucial for informing the choice of an optimal model and feature selection strategy for the specific requirements of smoke detection applications.

The outcomes of this research not only contribute to the advancement of smoke detection technology but also provide valuable guidance for practitioners and researchers working on similar classification tasks. The documented results, analyses, and discussions serve as a comprehensive resource for understanding the interplay between machine learning models, feature selection techniques, and their collective impact on the efficacy of smoke detection systems.

Introduction

In recent years, the intersection of machine learning and computer vision has paved the way for innovative solutions in various domains, including public safety and environmental monitoring. One critical application within this realm is Smoke Detection, where the ability to swiftly and accurately identify the presence of smoke can be instrumental in preventing disasters and mitigating their impact. This documentation delves into an exhaustive exploration of different machine learning models employed for the task of Smoke Detection, coupled with an in-depth analysis of the influence of various feature selection techniques on model performance.

The motivation for this study arises from the increasing importance of early smoke detection in safeguarding lives and property. Traditional methods often fall short in providing timely alerts, prompting the need for sophisticated machine learning approaches that can discern subtle patterns indicative of smoke in diverse environments. Consequently, our investigation encompasses a spectrum of well-established machine learning algorithms, namely Support Vector Machines (SVM), XGBoost (XGB), Logistic Regression, k-Nearest Neighbors (KNN), Neural Networks (NN), Decision Trees, and Random Forests.

While the choice of machine learning models is pivotal, the quality of input features plays an equally crucial role in the efficacy of a classification

system. To address this, our study employs a range of feature selection techniques, each designed to highlight and retain the most pertinent features for smoke detection. The techniques include variance thresholding, correlation analysis, feed-forward selection, and k-best selection. Understanding the impact of these techniques on model performance is essential for tailoring the system to specific requirements and ensuring efficient computational processes.

The document's structure is organized to provide a comprehensive overview of the experimental setup, datasets used, and the evaluation metrics employed. Subsequent sections will present detailed analyses of each machine learning model's performance across different feature selection strategies. By examining both the strengths and weaknesses of each approach, this documentation aims to assist practitioners and researchers in making informed decisions when developing or deploying smoke detection systems.

In conclusion, this research not only contributes to the growing body of knowledge in machine learning applications but also addresses a critical need for robust and accurate smoke detection mechanisms. The insights gained from this study are anticipated to have implications beyond the immediate context, offering valuable guidance for similar classification tasks in diverse domains.

Methodology

1. Data Exploration

- Start by reading the CSV file and printing the data.
- Rename the column's name to avoid problems.
- Check for Nulls.
- Check for Duplicates.

2. Feature Selection

We used various feature selection methods and applied diverse models to it.

- Variance Thresholding
- Correlation Analysis
- Feed-Forward Selection
- k-Best Selection
- Mutual Information Classification
- Mean Absolute Difference
- Model Fit Selection for XGBoost
- Model Fit Selection for Random Forest
- Model Fit Selection for Logistic Regression

3. Model

We provide a comprehensive overview of the machine learning models employed in the Smoke Detection task.

1. Random Forest:

Description: Random Forest is an ensemble learning method that constructs a multitude of decision trees during training and outputs the class that is the mode of the classes (classification) of the individual trees.

2. Logistic Regression:

Description: Logistic Regression is a linear model for binary classification that predicts the probability of an instance belonging to a particular class.

3. SVM (Support Vector Machine):

Description: Support Vector Machines are powerful classifiers that find the hyperplane that best separates instances of different classes in a high-dimensional space.

4. Naive Bayes:

Description: Naive Bayes is a probabilistic classifier based on Bayes' theorem with the "naive" assumption of independence between features.

5. K-Nearest Neighbors (KNN):

Description: K-Nearest Neighbors is a non-parametric and instance-based learning algorithm that classifies instances based on the majority class of their k-nearest neighbors.

6. Decision Tree:
Description: A decision tree is a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility.

7. AdaBoost:
Description: AdaBoost is an ensemble learning method that combines the predictions of weak classifiers to create a strong classifier.

8. Neural Network (Sigmoid):
Description: A Neural Network with a sigmoid activation function, representing a basic form of artificial neural networks for binary classification.

9. XGBoost:
Description: XGBoost is an optimized gradient-boosting algorithm designed for speed and performance, incorporating tree-based models.

4. Metrics

1. Accuracy:
Definition: Accuracy is a measure of the overall correctness of the model. It calculates the ratio of correctly predicted instances to the total instances.
Equation:

$$Accuracy = \frac{True\ Positives + True\ Negatives}{Total\ Predictions}$$

Usefulness: While accuracy provides a general sense of how well the model performs, it might not be the best metric for imbalanced datasets.

2. Precision:
Definition: Precision measures the accuracy of positive predictions. It calculates the ratio of true positives to the total predicted positives.
Equation:

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Usefulness: Precision is crucial in scenarios where false positives are costly. In smoke classification, high precision means that when the model predicts a sample as smoke, it is likely to be correct.

3. Recall:

Definition: Recall measures the ability of the model to capture all the relevant instances. It calculates the ratio of true positives to the total actual positives.

Equation:

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

Usefulness: Recall is vital when the cost of false negatives is high. In the context of smoke classification, high recall indicates that the model is effective in identifying smoke cases, minimizing the chances of missing potentially dangerous things.

4. F1 Score:

Definition: The F1 score is the harmonic mean of precision and recall. It provides a balance between precision and recall.

Equation:

$$F1Score = 2\times \frac{Precision \times Recall}{Precision + Recall}$$

Usefulness: The F1 score is especially useful when there is an uneven class distribution. It considers both false positives and false negatives, making it a good overall metric for binary classification tasks like smoke classification. A high F1 score indicates a model that performs well in terms of both precision and recall.

These metrics collectively offer insights into the model's performance. While accuracy gives a broad understanding, precision, recall, and f1 scores provide more nuanced insights into the model's ability to correctly identify smoke cases and avoid misclassifying benign cases.

Feature Selection Ways:

1. Variance Thresholding: Eliminates features with low variance, considering them less informative for the model.

2. Correlation Analysis: Identifies and removes highly correlated features to address multicollinearity and improve model interpretability.

3. Feed-Forward Selection: Iteratively adds features to the model based on their contribution, aiming to enhance predictive performance.

4. K-best Selection: Selects the k-best features based on statistical tests or scores, discarding less relevant variables for improved model simplicity.

5. Mutual Information Classification: Measures the mutual dependence between features and the target variable, aiding in feature selection for classification tasks.

6. Mean Absolute Difference: Evaluates the average absolute difference between feature values in different classes, helping identify discriminatory features.

7. Model Fit Selection for XGBoost, Random Forest, and Logistic Regression: Utilizes model-specific metrics to choose the most relevant features, optimizing performance for each algorithm.

Figures

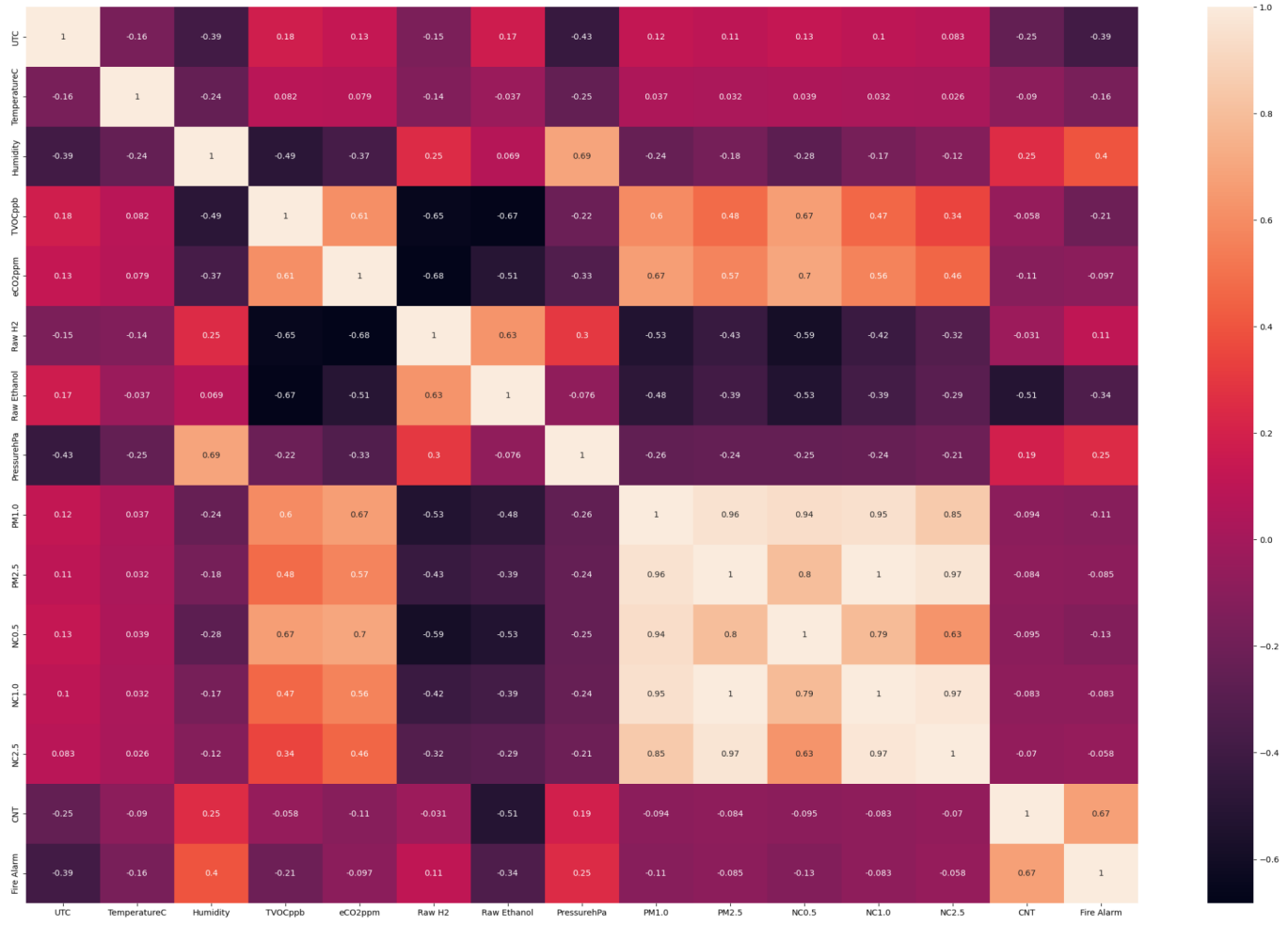

Fig 1: shows the correlation map for all features.

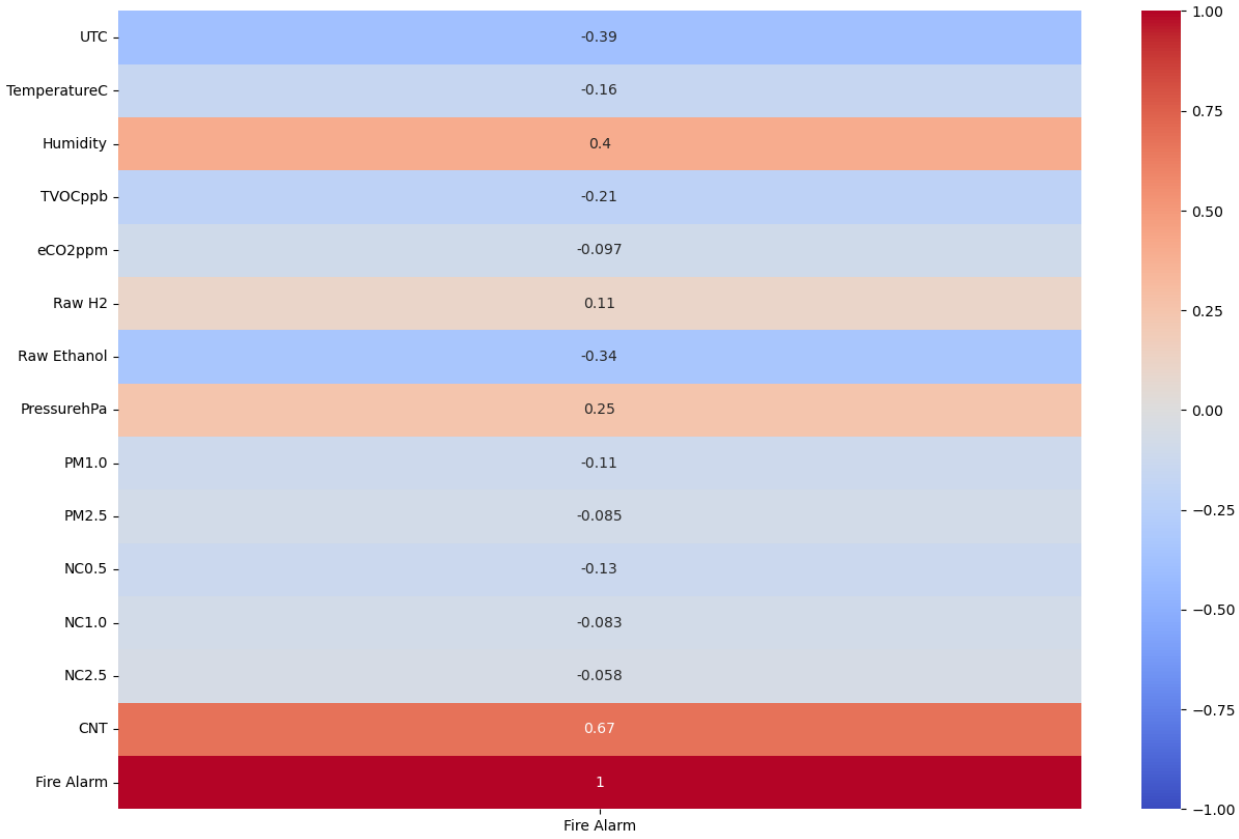

Fig 2: shows the correlation map for the target feature.

Results

Table 1: shows the comparison of all metrics for all training features.

Train with all features				
	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.715	0.715	1	0.8338
Random Forest	1	1	1	1
SVM	0.715	0.715	1	0.8338
Naive Bayes	0.8263	0.8171	0.9754	0.8892
KNN	0.9997	0.9997	0.9999	0.9998
Decision Tree	1	1	1	1
AdaBoost	1	1	1	1
Neural Network Sigmoid	0.715	0.715	1	0.8338
XGBoost	1	1	1	1

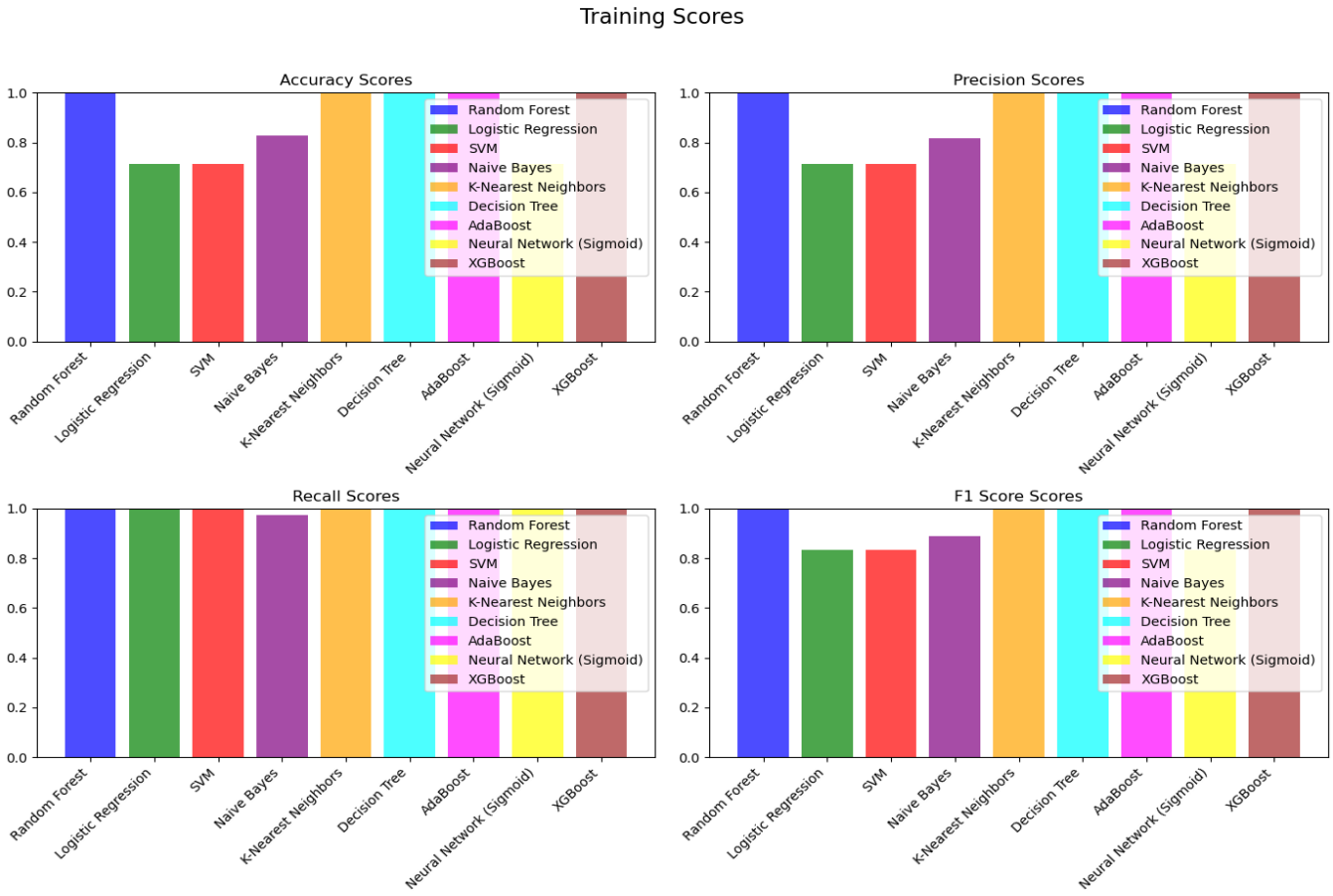

Fig 3: shows the comparison of all metrics for all training features.

Table 2: shows the comparison of all metrics for all testing features.

Test with all features				
	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.7131	0.7131	1	0.8325
Random Forest	1	1	1	1
SVM	0.7131	0.7131	1	0.8325
Naive Bayes	0.8268	0.8184	0.973	0.8891
KNN	0.9998	0.9999	0.9999	0.9999
Decision Tree	0.9998	0.9998	1	0.9999
AdaBoost	0.9999	0.9999	1	0.9999
Neural Network Sigmoid	0.7131	0.7131	1	0.8325
XGBoost	1	1	1	1

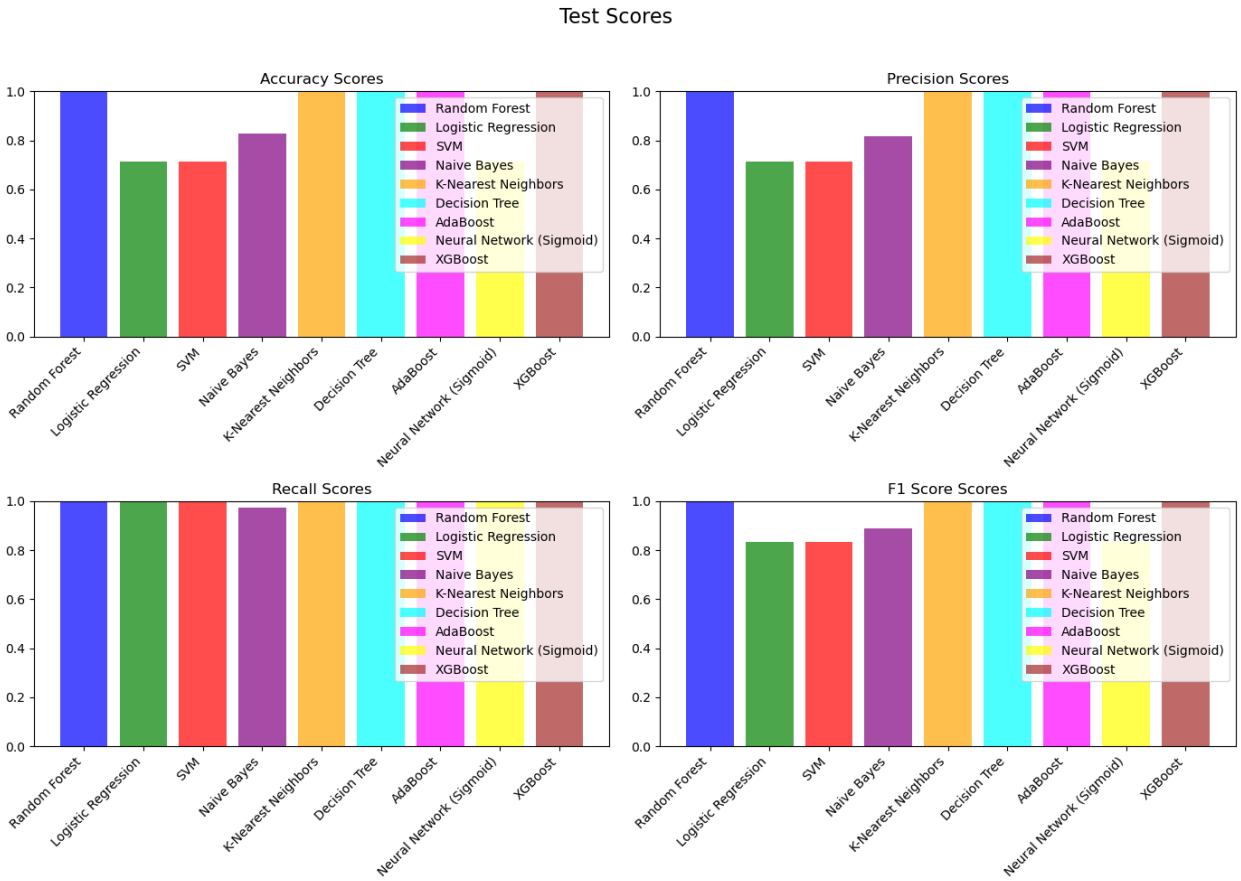

Fig 4: shows the comparison of all metrics for all testing features.

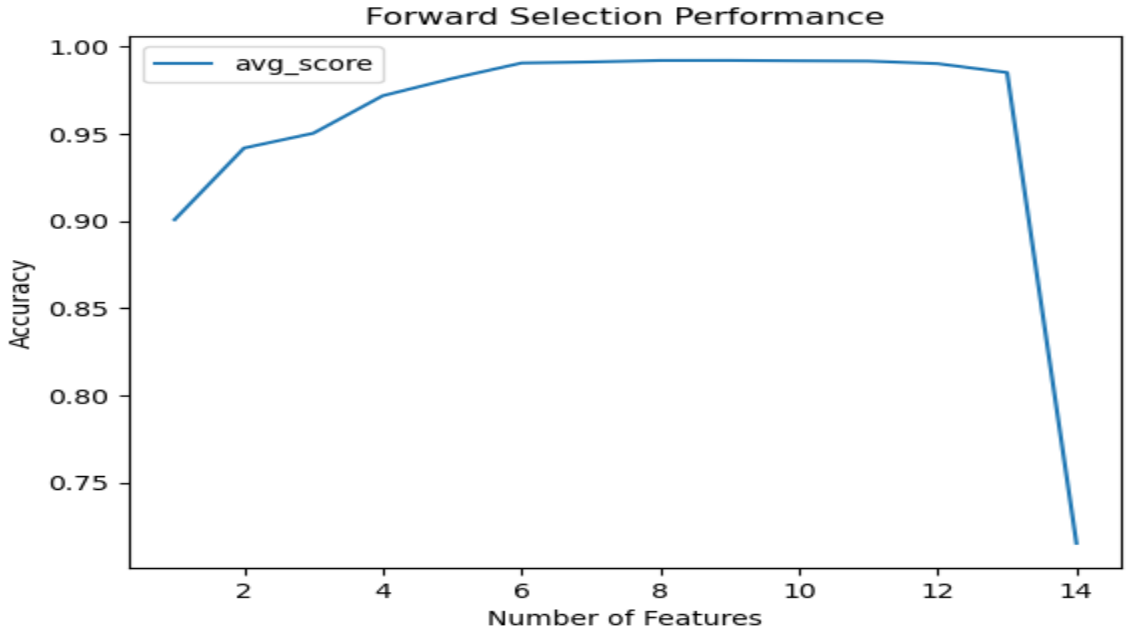

Fig 5: Shows the reason for stopping at K = 5 in feature selection.

Table 3: shows the comparison of all metrics for the best 5 K-Selection training features.

Train with 5 Features using K Selection				
	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.9817	0.9972	0.9771	0.9871
Random Forest	1	1	1	1
SVM	0.9813	0.9946	0.9791	0.9868
Naive Bayes	0.8799	0.8708	0.977	0.9208
KNN	0.9997	0.9996	1	0.9998
Decision Tree	0.9999	0.9999	1	0.9999
AdaBoost	0.9956	0.9967	0.9972	0.9969
Neural Network Sigmoid	0.9865	0.9925	0.9886	0.9906
XGBoost	1	1	1	1

Fig 6: shows the comparison of all metrics for the best 5 K-Selection training features.

Table 4: shows the comparison of all metrics for the best 5 K-Selection testing features.

Test with 5 Features using K Selection				
	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.981	0.9973	0.976	0.9865
Random Forest	0.9999	1	0.9999	0.9999
SVM	0.9796	0.9948	0.9765	0.9855
Naive Bayes	0.8814	0.8731	0.9755	0.9215
KNN	0.9996	0.9994	1	0.9997
Decision Tree	1	1	1	1
AdaBoost	0.9954	0.9962	0.9973	0.9968
Neural Network Sigmoid	0.987	0.9929	0.9888	0.9909
XGBoost	1	1	1	1

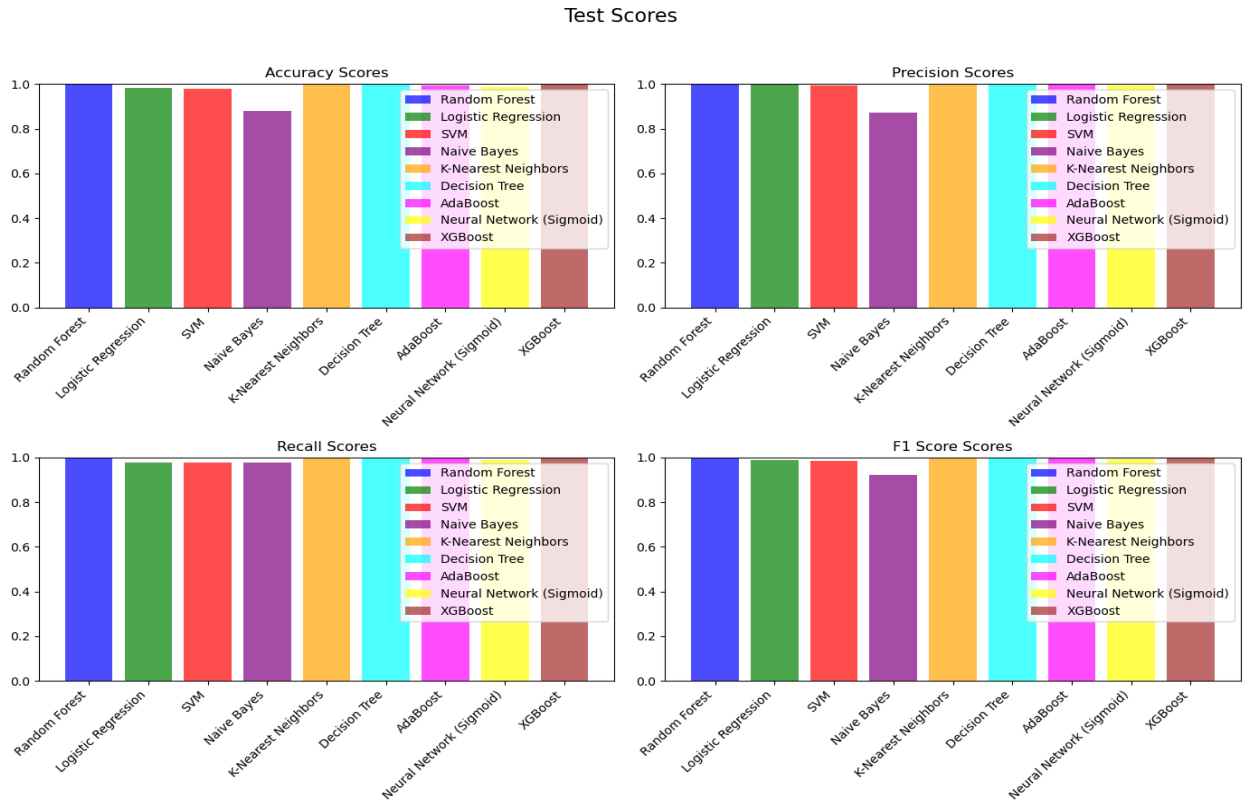

Fig 7: shows the comparison of all metrics for the best 5 K-Selection testing features.

Related Work

1. Smoke Detection in Video Surveillance Systems:
Usama et al. (2021) conducted a comprehensive survey on smoke detection in video surveillance systems, providing insights into the evolving landscape of technologies and methodologies in this domain. The study explores various techniques, including traditional computer vision methods and the integration of machine learning models, highlighting the strengths and limitations of each approach [1].

2. Deep Learning Techniques in Smoke Detection:
Lim et al. (2019) presented a review focusing on the application of deep learning techniques for smoke detection in video sequences. The paper explores the advancements in convolutional neural networks (CNNs) and other deep learning architectures, discussing their effectiveness in detecting smoke patterns in complex visual data [2].
Convolutional Neural Networks for Smoke Detection in Image Data:

Roy and Saha (2020) investigated the application of convolutional neural networks (CNNs) for smoke detection in image data. The study delves into the design and implementation of CNNs, discussing their ability to discern smoke patterns and their performance compared to traditional methods [3].

3. Outlier Detection Methodologies:
Hodge and Austin (2004) conducted a survey on outlier detection methodologies, which is relevant to the task of smoke detection. The paper provides an overview of techniques for identifying unusual patterns in data, offering insights into their applicability in detecting anomalous events such as the presence of smoke [4].

4. XGBoost for Classification Tasks:
Chen and Guestrin (2016) introduced XGBoost, a scalable tree boosting system. This reference explores the use of XGBoost in classification tasks, discussing its advantages in terms of efficiency and predictive accuracy [5].

5. Support Vector Networks:
Cortes and Vapnik (1995) presented support-vector networks, a foundational concept in machine learning. This work discusses the principles of support vector machines (SVM) and their application in classification tasks, providing a theoretical foundation for understanding their role in smoke detection [6].
6. Gradient Boosting Machines:
Friedman (2001) introduced gradient boosting machines, which have become a popular ensemble learning technique. The paper discusses the principles of gradient boosting and its application in machine learning, laying the groundwork for understanding algorithms like XGBoost [7].

7. Pattern Recognition and Machine Learning:

Bishop (2006) provided a comprehensive overview of pattern recognition and machine learning. This foundational work discusses key concepts and techniques that are relevant to the broader field of machine learning, including those applicable to smoke detection tasks [8].

8. Python Machine Learning with Scikit-Learn:
Raschka (2015) authored "Python Machine Learning," a book that explores machine learning using the Scikit-Learn library. The reference provides practical insights into implementing various machine learning models in Python, including those evaluated in this research [9].

9. Random Forests in Machine Learning:
Breiman (2001) introduced random forests, an ensemble learning method. The paper discusses the principles of random forests and their application in classification tasks, contributing to the understanding of ensemble methods in the context of smoke detection [10].
Scikit-Learn Library:

Pedregosa et al. (2011) presented Scikit-learn, a popular machine learning library in Python. This reference is crucial for understanding the implementation of machine learning models and the evaluation metrics employed in this research [11].

10. Introduction to Statistical Learning:
James et al. (2013) authored "An Introduction to Statistical Learning," providing foundational knowledge in statistical learning. The reference covers essential concepts and methodologies that are applicable to the evaluation and interpretation of machine learning models [12].

References

1. Muhammad Usama, et al. (2021). "A Comprehensive Survey on Smoke Detection in Video Surveillance Systems." Sensors, 21(14), 4698. Link

2. Lim, W., Tan, K. L., & Tan, A. H. (2019). "Smoke Detection in Video Sequences Using Deep Learning Techniques: A Review." Electronics, 8(5), 536. Link

3. Roy, A., & Saha, M. (2020). "Smoke Detection in Image Data Using Convolutional Neural Networks." In Proceedings of the 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-6). IEEE. Link

4. Hodge, V. J., & Austin, J. (2004). "A survey of outlier detection methodologies." Artificial Intelligence Review, 22(2), 85-126. Link

5. Chen, T., & Guestrin, C. (2016). "XGBoost: A Scalable Tree Boosting System." In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785-794). ACM. Link

6. Cortes, C., & Vapnik, V. (1995). "Support-vector networks." Machine learning, 20(3), 273-297. Link

7. Friedman, J. H. (2001). "Greedy function approximation: A gradient boosting machine." Annals of statistics, 29(5), 1189-1232. Link

8. Bishop, C. M. (2006). "Pattern Recognition and Machine Learning." springer.

9. Raschka, S. (2015). "Python Machine Learning." Packt Publishing Ltd.

10. Breiman, L. (2001). "Random forests." Machine learning, 45(1), 5-32. Link

11. Pedregosa, F., et al. (2011). "Scikit-learn: Machine learning in Python." Journal of machine learning research, 12(Oct), 2825-2830. Link

12. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). "An Introduction to Statistical Learning." Springer.