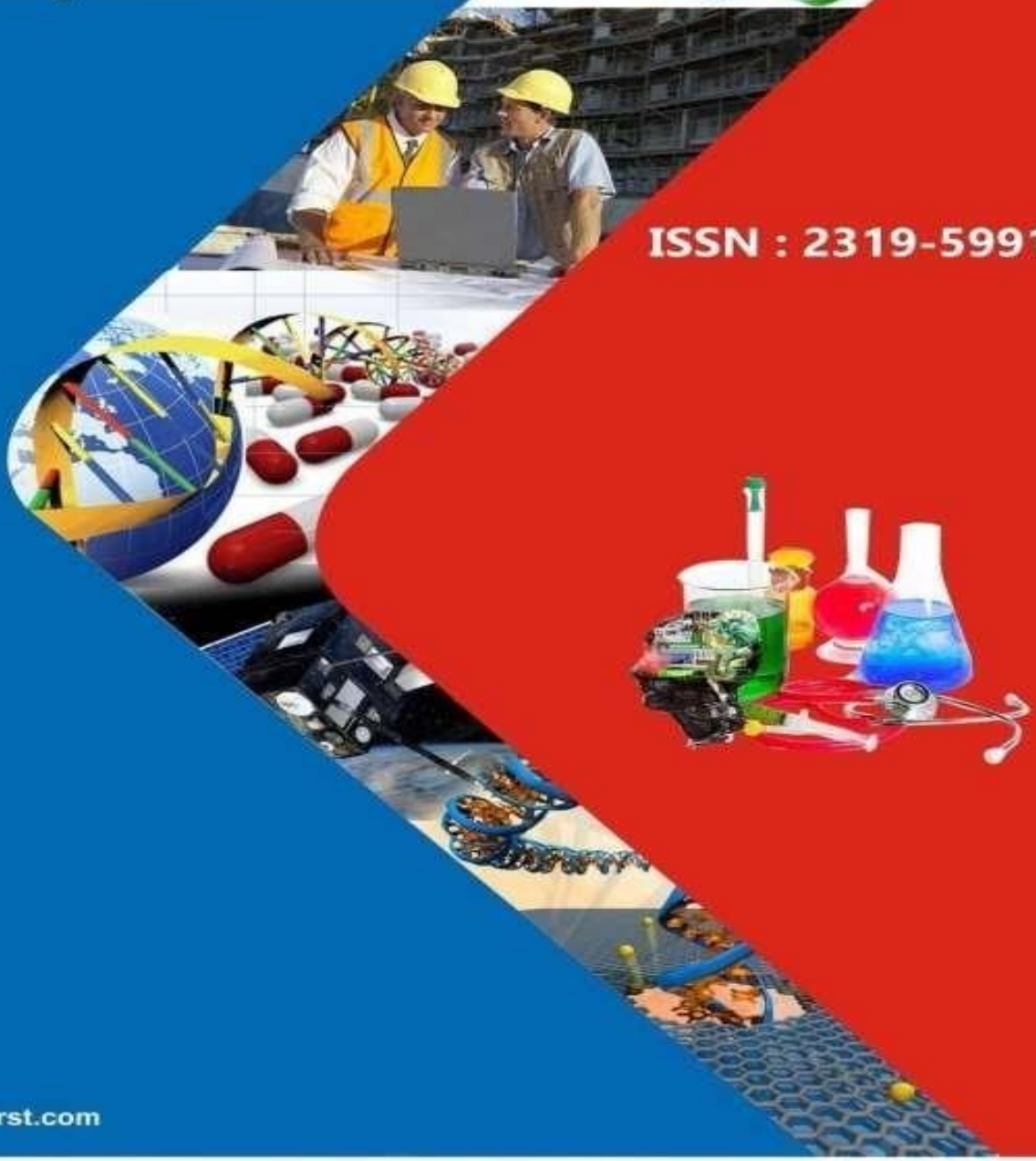
International Journal of Engineering Research and Science & Technology

IJERST

ISSN : 2319-5991



www.ijerst.com

Email: editor@ijerst.com or editor.ijerst@gmail.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

Use of Machine Learning Algorithms for the Prediction of Abnormal Behavior in Humans

, [1] Dr.Kamel Alikhan Siddique, [2]Munawar Hussain, [3] Mohammed Saif Pasha, [4] Syed Abdul Rahman
[1]Associate Professor,Department of CSE-AIML, Lords Institute of Engineering & Technology
[234]Student Department of CSE-AIML, Lords Institute of Engineering & Technology.

Abstract—

As an example, smoking at a petrol station is obviously not a good idea and might put people's lives in jeopardy. The goal of this research is to identify the most effective Machine Learning method for dealing with such prediction issues. Smoking, calling, and typical behaviors make up the datasets gathered for behavior detection. Linear Support Vector Machine (LSVM), Kernel Support Vector Machine (KSVM), Decision Tree Classifier (DT), Random Forest Classifier (RF), K-nearest Neighbors (KNN), and K-Means Clustering are some of the well-known algorithms studied in these experiments. Mean Squared Error (MSE) and Confusion Matrix are also used to evaluate the efficiency of each method. Lastly, PCA (Principal Component Analysis) shows how the optimal technique worked. With an accuracy of 82%, Random Forest Classifier (RF) outperforms all other methods and can predict when individuals will act in an aberrant way. Component-wise keywords: ML, dimensionality reduction, abnormal behavior prediction.

INTRODUCTION

Many harmful habits continue, despite the fact that more people are considering how their health affects their everyday lives. They are quite dangerous in some situations. Multitasking, such as talking on the phone and chatting at the same time, increases the likelihood of accidents, especially for drivers. Some public places, such as department stores and gas stations, have smoking bans due to the significant fire and explosion risks. Governments have enacted various restrictions to guarantee that individuals' wrong acts are discovered and dealt swiftly, as many lives may be spared if people would just stop from certain things. Still, it's impossible for humans to keep track of all these deeds. Lucky for us, advances in computer vision and machine learning that are easy on humans are making waves. Analyzing data linkages might teach computers to automatically classify photographs. As a result, images of smoking and talking on the phone may help teach computers to identify undesirable behavior.
In recent years, machine learning has made tremendous advances in various fields [1-3]. The integration of machine learning with human computer vision settings has been the subject of many previous studies. The computer was designed with a convolutional neural network that enables it to distinguish between distinct human movements [4]. Also, to keep an eye on how students were doing on the exam, Zhu et al. distributed a deep learning algorithm [5].

For the purpose of detecting smoking habits, Zhang et al. [6] have developed a machine learning system based on decision trees. With an accuracy rate of 84.11%, their model was the most effective.
Predicting people's calling habits using algorithms based on Machine Learning, in particular, is a relatively new field of study. Even without special glasses, it's hard to tell whether someone is smoking or talking on the phone. The fact that most people's hands aren't big enough to comfortably hold the phone is another potential issue. Using Support Vector Machine (SVM) and Convolutional Neural Network (CNN) based Machine Learning approaches, Zheng achieved a maximum accuracy of 93.5% in [7] when predicting how persons would move upstairs and downstairs. However, by comparing the most common machine learning algorithms, this study aims to discover the best effective approach for recognizing smoking and phoning behaviors.
The rest of this paper is organized as follows: The second part of the report describes the datasets and methods utilized to overcome the obstacles. Part 3 will follow, when we'll display the results of several machine learning approaches together with their respective classes. We will conclude the debate in Part 4.

METHOD

Section A. Defining and Preparing the Dataset Smoking, Calling, and Normal are the three classes that make up the dataset that this article utilizes. This article utilizes the 'Cigarette Smoker Detection' dataset from Kaggle, which has

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

805 photos of varying sizes, for the Smoking class [8]. There are 1,227 photos from the TIANCHI DATA SET with varying sizes and 396 images from the CSDN with dimensions of 3456 × 4608 for the Calling class [9, 10]. The Normal class is derived from the 'Person Face Dataset' on Kaggle, which includes 10,000 1024×1024 photos [11]. Images from the sample are shown in Figures 1, 2, and 3.



Figure 1. Sample images in Smoking class



Figure 2. Images in Calling class from TIANCHI DATA SET



Figure 3. Sample images in Calling class from CSDN

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

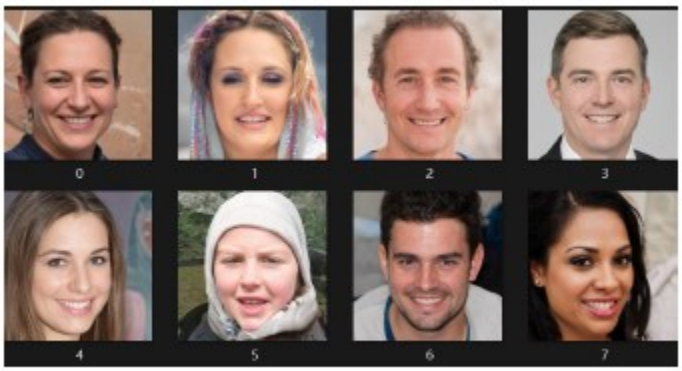


Figure 4. Sample images in Normal class

There are six components to the preprocessing. The process begins with locating the human face in photos using the get_frontal_face_detector function from dlib. This will allow for more accurate detection of their actions, such as phone calls. The photos are then downsized to a 64×64 format. Part three involves using the cvtColor function from cv2 to turn pictures into grayscale. This makes them more consistent for the machine learning system. Then, as the Calling class only has around 700 photos, 700 preprocessed images are chosen for each class to ensure dataset parity. Also, the dataset is divided by 255 to make it more typical in this study. The dataset is partitioned into train and test sets with a training-to-test ratio of 0.8. The processed data is shown in Figures 5, 6, and 7.

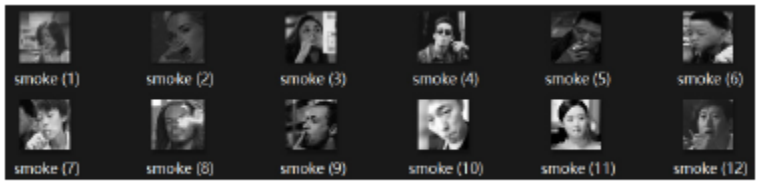


Figure 5. Preprocessed images in Smoking class

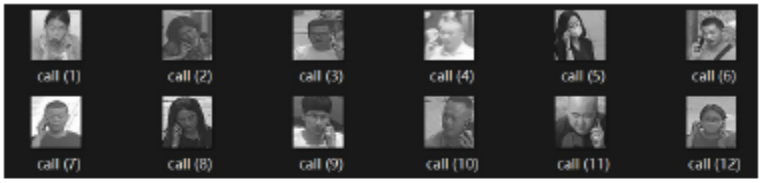


Figure 6. Preprocessed images in Calling class

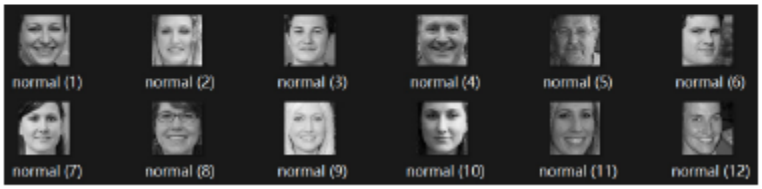


Figure 7. Preprocessed images in Normal class

Section B: Learnt Algorithms Popular Machine Learning methods such as Support Vector Machine, Decision Tree, Random Forest, K-Nearest Neighbors, and K-Means were used in this study. Listed below are a few introductions to these algorithms. a) SupportVectorMachine (SVM): SVM is a model for supervised machine learning that can handle regression and classification issues. Its goal is to find a hyperplane in an N-dimensional space that can categorize the points of data. There is a plethora of hyperplanes from which to choose when separating two data points from classes. In order to improve its classification accuracy, support vector machines (SVMs) seek for hyperplanes with the largest distances between data points. There are several different kernels available in SVM,

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

including linear, poly, rbf, sigmoid, and precomputed. Regularization parameter (C) and kernel coefficient (Gamma) for 'rbf', 'poly', and'sigmoid' in support vector machines (SVM). In order to assess the models' capabilities in categorizing the dataset, this article utilizes LSVM—'linear' kernel and KSVM—'rbf' kernel. Parameters C and Gamma are default values of 1 and "scale" respectively. b) Decision Tree: If you're having trouble with classification or regression, you may utilize this non-parametric supervised learning approach. Decisions and their potential outcomes are shown by means of a tree-like representation. The test results are represented by the branches, and the class labels are represented by the leaf nodes. c) Random Forest: This approach constructs a network of decision trees to address classification or regression issues. The random forest returns the class that the majority of trees choose while doing a classification job. The average forecast from each tree is the output for regression tasks. To prevent decision trees from becoming overfit to their training data, Random Forest is a good option. The integer 'n_estimators,' representing the total number of trees in the Random Forest, is its primary parameter. In this work, the issue is solved using a model with 'n_estimators' set at 250.

III. RESULT ANDDISCUSSION

Tables I, II, III, IV, V, and VI show the confusion matrix and outcomes of the various algorithms, including KSVM, LSVM, Decision Tree, Random Forest, KNN, and K-means.

TABLE I. CONFUSION MATRIX OF KSVM

CONFUSION MATRIX OF KSVM			
Predict / Actual	*Smoking*	*Calling*	*Normal*
Smoking	87	40	5
Calling	66	87	4
Normal	6	3	128

TABLE II. CONFUSION MATRIX OF LSVM

CONFUSION MATRIX OF LSVM			
Predict / Actual	*Smoking*	*Calling*	*Normal*
Smoking	105	25	2
Calling	59	91	7
Normal	2	1	134

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

TABLE III. CONFUSION MATRIX OF DECISION TREE

CONFUSION MATRIX OF DECISION TREE			
Predict / Actual	Smoking	Calling	Normal
Smoking	94	31	7
Calling	48	92	17
Normal	15	6	116

TABLE IV. CONFUSION MATRIX OF RANDOM FOREST

CONFUSION MATRIX OF RANDOM FOREST			
Predict / Actual	Smoking	Calling	Normal
Smoking	97	33	2
Calling	32	122	3
Normal	5	2	130

TABLE V. CONFUSION MATRIX OF KNN

CONFUSION MATRIX OF KNN			
Predict / Actual	Smoking	Calling	Normal
Smoking	93	9	30
Calling	78	45	34
Normal	2	0	135

TABLE VI. CONFUSION MATRIX OF K-MEANS

CONFUSION MATRIX OF K-Means			
Predict / Actual	Smoking	Calling	Normal
Smoking	33	67	32
Calling	38	74	45
Normal	109	3	25

TABLE VII. RESULTS OF DIFFERENT ALGORITHMS

RESULTS OF DIFFERENT ALGORITHMS					
	Accuracy	MSE	Precision	Recall	F1-score
KSVM	0.71	0.369	0.72	0.71	0.71
LSVM	0.77	0.254	0.78	0.77	0.77
Decision Tree	0.71	0.446	0.71	0.71	0.71
Random Forest	0.82	0.230	0.82	0.82	0.82
KNN	0.64	0.585	0.69	0.64	0.60
K-Means	0.31	1.683	0.33	0.31	0.31

Accuracy Comparison The findings in Table VII demonstrated that Random Forest, Linear Support Vector Machine, Decision Tree, and Kernel Support Vector Machine performed admirably in terms of accuracy. The most effective of them is Random Forest, which gets 82% of the results right. K-means is the poorest, with an accuracy of just 31%, whereas KNN is not terrible either. Mean Squared Error Comparison Among the models tested, Linear Support Vector Machine (LSVM), Kernel Support Vector Machine (KSN), Decision Tree (DT), Random Forest (RF), and KNN all shown strong performance, with LSVM achieving the best overall result (MSE=0.254). In comparison, K-Means clustering performs horribly with an MSE of 1.683.

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

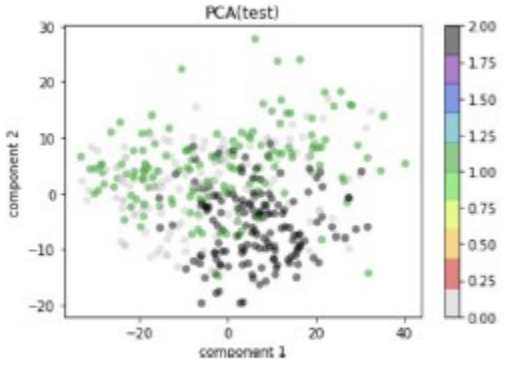


Figure 8. Visualization of the relationship between test_x and test_y.

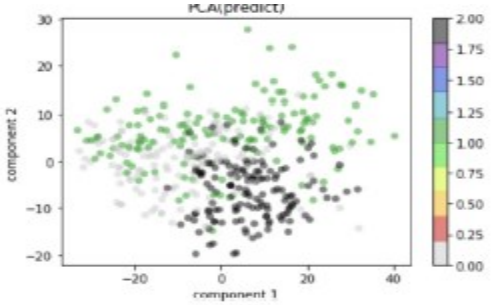


Figure 9. Visualization of the relationship between test_x and predicted_y by Random Forest.

CONCLUSION

The research aims to use Machine Learning algorithms to identify individuals whose actions might endanger the lives of others. Machine learning algorithms such as Decision Tree, K-Means Clustering, K-Nearest Neighbors (KNN), Decision Tree, Kernel Support Vector Machine (KSVM), and Linear Support Vector Machine (LSVM) are the main focus of this paper. To determine whether the model is good or not, tools like the Confusion Matrix and Mean Squared Error are used. The best algorithm's output may also be seen using Principal Component Analysis. Random Forest is the best solution for this issue, according to the study's findings, but K-Means clustering is the worst. Soon, there will be apps that can use cameras to identify people's actions, and by tweaking the settings, the Random Forest model will perform even better.

REFERENCES

[1] Q.Zhou,W.Lan,Y.ZhouandG.Mo,"EffectivenessEvaluationofAnti-bird Devices based on Random Forest Algorithm," 2020 7thInternational Conference on Information, Cybernetics, andComputational Social Systems (ICCSS), 2020, pp. 743-748.

[2] Y. Guo, Y. Zhou, X. Hu and W. Cheng, "Research on Recommendationof Insurance Products Based on Random Forest," 2019 InternationalConference on Machine Learning, Big Data and Business Intelligence(MLBDBI), 2019, pp. 308-311.

[3] Y. Qiu, P. Chen, Z. Lin, Y. Yang, L. Zeng and Y. Fan, "ClusteringAnalysis for Silent Telecom Customers Based on K-means++," 2020IEEE 4th Information Technology, Networking, Electronic andAutomation Control Conference (ITNEC), 2020, pp. 1023-1027.

[4] S. Zheng, “Research on Algorithm of Pedestrain Attitude Estimation andRecognitionBasedonMachineLearning,”ShandongUniversity,2019,.DOI:10.27272/d.cnki.gshdu.2019.000463.

[5] G. Zhu, X. Jiang, F. Xu, “Application of video behavior and actionrecognition based on machine learning in

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com

Vol. 21, Issue 2, 2025

paperless assessment (inChinese),” Construction informatization in China, 2019, vol. 10, pp. 56-57.

[6] Y. Zhang, J. Liu, Z. Zhang and J. Huang. "Prediction of daily smokingbehavior based on decision tree machine learning algorithm." In 2019IEEE 9th International Conference on Electronics Information andEmergency Communication (ICEIEC), pp. 330-333. IEEE, 2019.

[7] Y. Zheng, “Research on Machine Learning Algorithm for HumanBehavior Recognition,” Wuhan University of Technology, 2019. TheDOI: 10.27381 /, dc nki. Gwlgu. 2019.000606.

[8] “Cigaretter-somker-detection”, kaggle,https://www.kaggle.com/datasets/vitaminc/cigarette-smoker-detection,2019.

[9] “Smoking and Calling Image Dataset”, Tianchi,https://tianchi.aliyun.com/dataset/dataDetail?dataId=89271,2022.

[10] “Calling dataset”, CSDN.https://download.csdn.net/download/weixin_45546050/77837488,2022.

[11] “Person-face-dataset”, kaggle,https://www.kaggle.com/datasets/almightyj/person-face-dataset-thispersondoesnotexist,2022.