COMPUTOLOGY: Journal of Applied Computer Science and Intelligent Technologies
Volume 5, Issue 2, Jul-Dec 2025, pp. 34-53
DOI: 10.17492/computology.v5i2.2503
https://www.journalpressindia.com/cjacsit

Development of an Improved Facial Analysis-based System for Predicting Drug Addiction using Random Forest Classification Algorithm

Ismail Akuji, Sulaiman Abdulsalam**, Ronke Babatunde*** and Olugbemi Olaniyan****

ABSTRACT

Drug abuse has become a widespread global issue, affecting individuals, families, and communities. Machine learning techniques have shown promise in combating drug addiction prevalence through early prediction and immediate intervention. This study proposes an improved facial analysis-based drug addiction prediction system using a random forest classification algorithm, trained on facial images. Feature extraction and selection were performed using a histogram of oriented gradients and recursive feature elimination, respectively. The random forest classification model was tuned with grid search cross-validation, and evaluated using accuracy, precision, recall, and F1-score. Tuning the system significantly improved its performance, with accuracy increasing from 84.62% to 87.18%, precision from 82.61% to 83.33%, recall from 90.48% to 95.24%, and F1-score from 86.37% to 88.89%. This increase demonstrates the importance of hyperparameter tuning and the robustness of the random forest algorithm. Future studies can improve upon this work by incorporating a larger facial dataset for better practical results.

Keywords: *Drug addiction; Random forest; Grid search cross-validation; Recursive feature elimination; Histogram of oriented gradient.*

1.0 Introduction

Drug addiction has become a global issue and its aftermath affects not only drug abusers but also families and communities worldwide due to its physical and mental health deterioration, strained relationships, economic instability, and even mortality.

**Corresponding author; Student, Computer Science Department, Kwara State University, Malete, Kwara, Nigeria (E-mail: akujiismaheel@gmail.com)*
***Senior Lecturer, Computer Science Department, Kwara State University, Malete, Kwara, Nigeria (E-mail: sulaiman.abdulsalam@kwasu.edu.ng)*
****Associate Professor, Computer Science Department, Kwara State University, Malete, Kwara, Nigeria (E-mail: ronke.babatunde@kwasu.edu.ng; olugbemi.olaniyan@kwasu.edu.ng)*

The classification of drugs is based on their effect on the users' bodies. Basically, medicinal and recreational drugs are the two types of drugs. Medicinal drugs are commonly used for treatment purposes, such as reducing pain in the body or facilitating treatment, whereas recreational drugs are drugs taken for enjoyment and are often referred to as psychoactive or illicit drugs and often lead to addiction. In Nigeria, the menace of drug use is prevalent, particularly among the youths. The United Nations Office on Drugs and Crime [UNODC] (2022) reports the engagement of approximately 14.3 million Nigerians in drug intake. However, efforts have been made to combat the persistence of drug abuse in the society, encompassing traditional and automated approaches.

Traditional approach to drug addiction prediction encompasses the use of non-pharmacological methods such as behavioral therapies, counseling, and support groups (Gu *et al.*, 2021; Basuni & Siregar, 2023), which are known for time consumption and inherent biases, leading to the emergence of the automated approach, such as machine learning, internet of things among others. IoT technology is revolutionizing healthcare through real-time monitoring, data-driven insights, and personalized patient care, leading to improved medical systems and outcomes (Chavan *et al.*, 2024).

This healthcare field leverages various machine learning techniques to identify patterns in patients' data and apply them to make informed predictions (Sharma *et al.*, 2023), leading to timely intervention and diagnosis. Recent studies have demonstrated the potential of machine learning algorithms such as Random Forest in predicting drug addiction (Parekh & Fahim, 2021; Oliva *et al.*, 2022). Yet, their effectiveness on image-based data remains underexplored, highlighting a significant gap in research on facial data-driven drug addiction prediction. Random forest is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy and reduce overfitting.

Feature selection plays a crucial role in improving the performance of machine learning model, particularly for dataset with many features (Lakshmi & Das, 2023) such as image dataset. Recursive Feature Elimination (RFE), however, is one of the wrapper feature selection techniques that has shown promise in hybridizing with Random Forest classifiers (Jeon & Oh, 2020) and handling high-dimensional datasets (Ramezan, 2022). More precisely, RFE is a backward selection approach (Mahmoud & Garko, 2022) whose process entails the continuous removal of the least important features until a threshold is reached, thereby retaining the most important features.

1.1 Objectives of the study

This study aims to:

- Collect and preprocess facial dataset to remove noise from the images.
- Perform feature extraction using histogram of oriented gradients.

- Perform feature selection using recursive feature elimination technique.
- Develop a drug addiction prediction system using random forest classification algorithm.
- Optimize the algorithm using grid search cross-validation.
- Evaluation the system using accuracy, precision, recall, and f1-score.

This study is significant to healthcare professionals, individuals vulnerable to drug addiction, researchers, and policymakers. The development of a facial analysis-based drug addiction prediction system can facilitate early intervention and treatment, ultimately mitigating the devastating consequences of drug abuse. Furthermore, the findings of this can contribute to the development of more accurate and efficient drug addiction prediction systems.

2.0 Literature Review

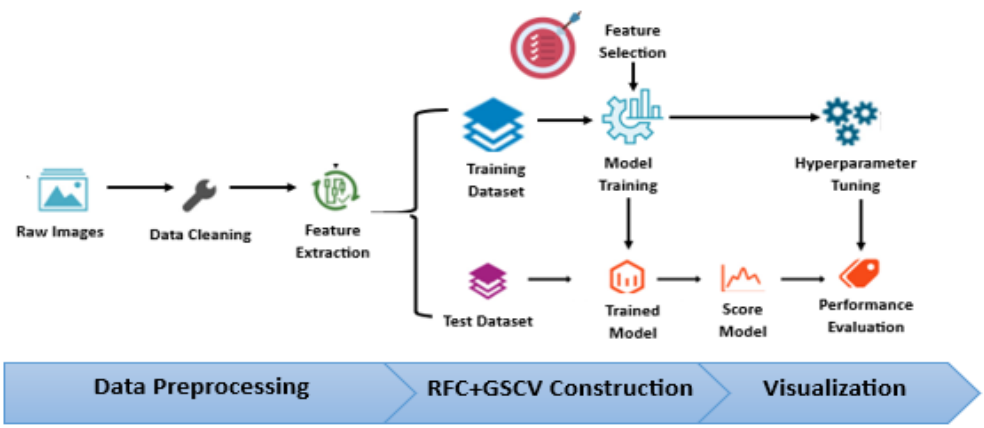
Existing related works on drug addiction prediction using machine learning have been understudied, and a few are discussed as follows. Parekh and Fahim (2021) assessed the efficiency of machine learning models in projecting marijuana intake and its associated factors using logistic regression, decision trees, random forest-Gini function, and naïve Bayes. The study found that random forest realized 97%, 96%, 94%, 93%, and 94% of AUC, accuracy, recall, precision, and F1-score scores, respectively, outperforming other selected algorithms. Another study (Haque *et al.*, 2021) employed random forest, extreme gradient boost, and Gaussian naïve Bayes to predict depression in childhood and adolescence. The study adopted data from the second Australian Child and Adolescent Survey of Mental Health and Wellbeing (2013-2014) and employed Boruta-RF for selecting important features. Random forest outperformed other algorithms in terms of accuracy (95%) and precision (99%). The study, however, failed to compare the result of the model before and after feature selection in order to juxtapose the effectiveness of Boruta-RF.

Choi *et al.* (2021) used logistic regression, random forest, and k-nearest neighbour to develop a predictive model for cannabis addiction using the 2019 National Survey on Drug Use and Health survey of 698 participants. The results show that random forest outperform other algorithms having achieved 99.8% accuracy scores. This demonstrates the efficacy of the random forest algorithm to predict cannabis. Gong *et al.* (2021) used the gradient boosting method (GBM) on 895 male substance users through the survey method. GBM identified 10 influencing factors, which include live events, deviant peers, and others, and findings show the potential of the method for the prevention of substance craving attitudes by users of drugs. However, the study was flawed owing to its restriction to a male sample. Furthermore, Uddin *et al.* (2022) applied sentiment analysis to measure the

effectiveness of drugs using naïve Bayes, random forest, support vector machines, and multilayer perceptron algorithms. The models were trained on data acquired from the UCI repository. Results prove RF as the best-performing classifier with an accuracy score of 94.06%. Oliva *et al.* (2022) identified factors of different substance use disorders (SUDs) in bipolar disorder (RD) using random forest. The study reveals that alcohol use disorder (AUD) could be found in bipolar disorder, having obtained a 75% score in each of the sensitivity and specificity scores. Conclusively, the model demonstrates a promising tool to foretell the risk of SUD in BD, though its effectiveness depends on socio-demographic or clinical factors.

Basuni and Siregar (2023) addressed the growing issue of drug abuse and addiction by employing artificial neural networks, decision trees, k-nearest neighbours, support vector machines, and random forests to develop models for the classification of users and non-users of drugs on the UCI repository dataset. Results confirmed that random forest outperformed other algorithms, having attained a 93% score in each of accuracy, precision, and recall; an 89% F1 score; and a 0.51 area under the curve. The study recommends the exploration of other evaluation metrics and large datasets for future studies. Almahmood *et al.* (2023) applied Gaussian naïve Bayes, logistic regression, k-nearest neighbour, random forest, and decision tree to classify drug users and non-users. The model is trained with the UCI repository dataset containing 18 illicit drugs. Out of the 18 drugs, random forest achieved the best accuracy in 9 drugs, ranging from 70% to 89%. Above all, machine learning algorithms such as Random Forest have been shown as effective algorithms for the development of drug addiction prediction models, particularly on numerical datasets. To generalize this effectiveness, however, it is imperative to explore facial images.

Figure 1: Framework of the Proposed System



Source: Adapted from Li et al. (2021)

3.0 Methodology

3.1 Project approach

From Figure 1, the developmental phases of the proposed drug addiction system are depicted which consists of data preprocessing, model construction, and visualization. As shown in Figure 1, this study adapted Li *et al.* (2021) framework with key modifications which includes direct feature extraction from the dataset; inclusion of feature selection to optimize input data quality; model's performance comparisons before and after hyperparameter tuning; hybrid of use of Random Forest Classifier (RFC) and Grid Search Cross-Validation (GSCV) for model training. These modifications aim to improve the system's efficiency and accuracy.

3.1 Dataset preprocessing

Data were collected through web scraping and survey methods. The formal method is employed to collect images of drug addicts only, whereas the latter method is applied to acquire the images of both categories, drug and non-drug addicts. 64 images were collected, spanning drug addicts and non-drug addicts. This study focuses on the important facial regions, shown in Table 1.

Table 1: Facial Area Points

Facial Area	Points
Left Eye	37-42
Right Eye	43-48
Mouth	49-68
Nose	28-36
Left Cheek	1-9
Right Cheek	10-17

Source: Amato et al. (2018)

Moreover, the collected data were pre-processed in terms of normalization, augmentation, transformation, and dimensionality reduction. Normalization ensures that the images are of equal dimension (128x128). Augmentation entails the process of increasing the size of the collected dataset via rescaling, shearing, zooming, and shifting, making the images increase from 64 to 192 images.

3.2 Feature extraction

Histogram of Oriented Gradient (HOG) is used for the feature extraction technique. The algorithm of image gradient calculation is presented in Figure 2.

Figure 2: Algorithm for Histogram of Oriented Gradients

- For function f(x,y), the gradient is the vector (f_x, f_y)
- At each pixel, image gradient horizontal (x-direction) is calculated by:

$$magnitude = \sqrt{(f_x^2 + f_y^2)}$$

- vertical (y-direction) is calculated by:

$$direction\ (\theta) = tan^{-1}\left(\frac{f_y}{f_y}\right)$$

- Split each into angular bins (9 bins with 0-180^0, 20^0 each bin)
- Create a histogram of generated gradient vectors.
- Make overlapping blocks from cells:

$$Number\ of\ blocks\ = \frac{(Image\ size\ -\ Block\ size)}{Stride} + 1$$

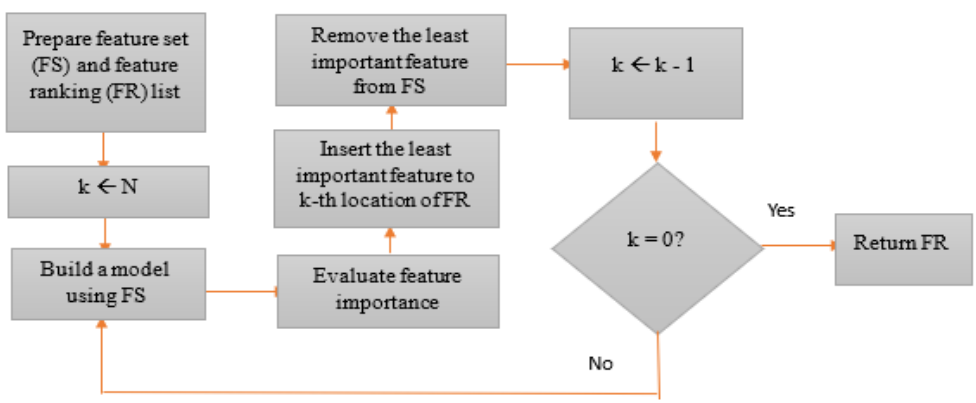
Where:
- Image size is 16*16 pixels
- Block size is 2*2 cell
- Stride is 1 overlapping step

- Calculating feature vector:

$$size\ of\ features\ extracted = size\ of\ horizontal\ blocks * size\ of\ vertical\ bocs * feature\ vector\ points$$

Where: - size of horizontal and vertical blocks = 15 - feature vector points = 36

Figure 3: Framework of RFE Process



Source: Jeon & Oh (2020)

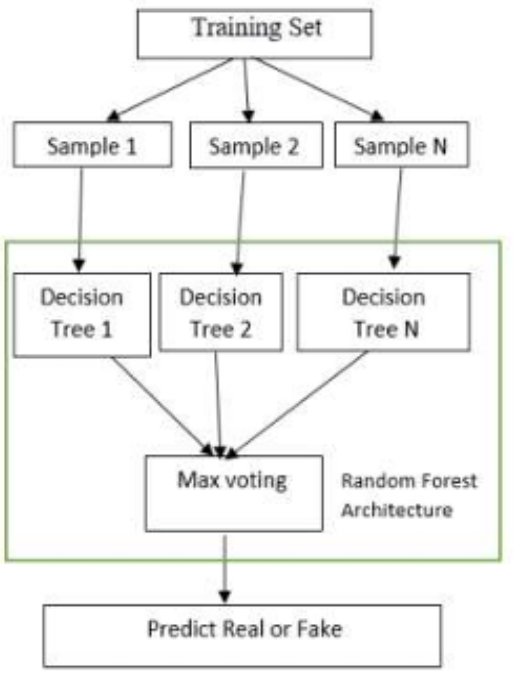
3.3 Feature selection

The rationale behind feature selection is to ensure that a model is trained with relevant features, rather than irrelevant ones. Figure 3 depicts the flowchart of the feature selection process, starting from feature ranking and ends with the return of the important features with highest ranking score.

3.4 Model development using RFC and grid search CV

The framework of the random forest algorithm is presented in Figure 4.

Figure 4: The framework of random forest



Source: Deedee et al. (2024)

The framework for the optimizing the RFC with Grid Search CV is presented in Figure 5.

Figure 5: RFC+GSCV Algorithm

Given full training set (x_k, y_k)

1. Perform Grid Search CV with k-fold cross-validation to tune the RFC model
2. Train the RFC model using important features:
3. Initialize T, n_k

4. Repeat T times
5. For (k=1 to n) Do
6. RFC_Model ← fit(x_k, y_k), F(n)
7. Compute F((n) importance scores)
8. While (importance score (n_k) < threshold)
9. feature.Drop(n_k)
10. Go back to 5
11. Else: Return the remaining F(n_k)
12. End

Source: Adapted from Misra & Yadav (2020) & Mahmoud & Garko, (2022)

3.5 Performance evaluation metrics

Accuracy measures the proportion of correct predictions out of the total number of predictions made. Mathematically, accuracy can be represented as (Choi *et al.*, 2021):

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad ...1$$

Precision determines the proportion of true positive predictions out of all positive predictions made. It can be calculated using the formula in Equation 3.2 (Choi *et al.*, 2021).

$$Precision = \frac{TP}{TP+FP} \quad ...2$$

Recall calculates the proportion of true positive predictions (correctly predicted as drug addicts) out of all actual positive outcomes. Equation 3.3 presents the mathematical calculation of Recall (Choi *et al.*, 2021).

$$Recall = \frac{TP}{TP+FN} \quad ...3$$

F1-Score represents a weighted mean of Equations 3.2 and 3.3. It can be mathematically measured using (Choi *et al.*, 2021):

$$F1 - Score = \frac{2*(Precision * Recall)}{Precision+Recall} \quad ...4$$

True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN), which denote correctly identified positive outputs, correctly identified negative outputs, negative outputs incorrectly identified as positive, and positive outputs incorrectly identified as negative, respectively.

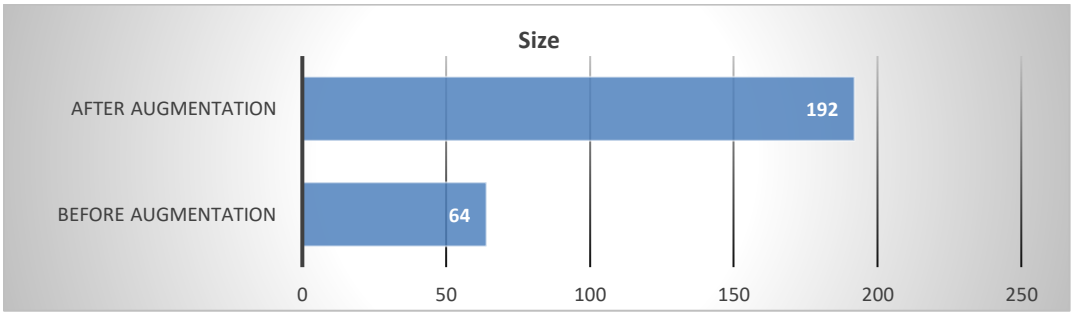
4.0 Results and Discussion

4.1 Result from dataset visualization

Figure 6 shows the visualized dataset size before and after augmentation. During the process of the feature extraction, images are divided into blocks, containing a 16x16 grid. Figure 7 shows the grids containing pixel values of an image. The blocks are divided

into 3x3 matrices, whereby the last row and column are left over due to the block size not being perfectly divisible by the matrix size.

Figure 6: Size of Dataset Before and After Augmentation



Source: Jupyter Notebook Code Editor

4.2 Results from Feature Extraction Process

Figure 7: Representation of an Image using 16x16 Grid Block

Matrix 1			Matrix 2			Matrix 3			Matrix 4			Matrix 5			Left-over
89	85	87	86	54	52	65	65	75	100	114	122	73	93	99	103
89	88	87	73	61	81	66	63	60	92	121	144	143	90	97	100
87	87	83	53	96	91	70	63	64	100	135	156	174	141	95	101
Matrix 6			**Matrix 7**			**Matrix 8**			**Matrix 9**			**Matrix 10**			**Left-over**
88	87	69	97	112	84	69	62	65	92	117	144	172	185	140	102
86	86	88	133	93	53	40	45	53	52	52	85	146	196	190	111
87	87	103	111	56	34	29	52	57	81	40	34	74	121	206	117
Matrix 11			**Matrix 12**			**Matrix 13**			**Matrix 14**			**Matrix 15**			**Left-over**
86	85	109	93	85	30	27	45	85	130	29	33	105	142	193	113
87	87	117	117	86	33	33	65	95	153	63	38	90	175	205	134
84	73	120	111	68	54	86	106	103	183	177	129	135	171	211	153
Matrix 16			**Matrix 17**			**Matrix 18**			**Matrix 19**			**Matrix 20**			**Left-over**
82	80	110	134	87	61	78	99	93	164	165	99	152	184	208	126
80	85	122	118	100	66	71	59	61	63	115	118	139	179	192	169
77	75	109	103	111	92	69	46	44	89	139	153	169	161	174	143
Matrix 21			**Matrix 22**			**Matrix 23**			**Matrix 24**			**Matrix 25**			**Left-over**
74	73	92	91	106	57	70	65	65	102	134	95	142	151	165	112
74	72	76	86	91	55	68	59	55	83	135	93	122	148	153	86
71	70	69	80	76	60	60	47	38	56	98	106	119	139	122	79
Left-over															
70	68	68	81	66	50	61	53	45	93	121	96	114	132	93	76

Source: Jupyter Notebook Code Editor

Table 2: Determination of the Image' Oriented Gradients

Matrix	X-direction (X-dir) and Y-direction (Y-dir)	Gradient Magnitude $\sqrt{(f_x^2 + f_y^2)}$	Gradient *Direction* (θ) $tan^{-1}\left(\frac{f_y}{f_y}\right)$
1	X-dir = 87-89 = -2 Y-dir = 85-87 = -2	$\sqrt{(-2)^2 + (-2)^2} = 2.8$	$tan^{-1}\left(\frac{-2}{-2}\right) = 45^0$
2	X-dir = 81-73= 8 Y-dir = 54-96 = -42	$\sqrt{(8)^2 + (-42)^2} = 42.8$	$tan^{-1}\left(\frac{-42}{8}\right) = -79.2^0$
3	X-dir = 60-66 = -6 Y-dir = 65-63 = 2	$\sqrt{(-6)^2 + (2)^2} = 6.3$	$tan^{-1}\left(\frac{2}{-6}\right) = -18.4^0$
4	X-dir = 144-92 = 52 Y-dir = 114-135 = -21	$\sqrt{(52)^2 + (-21)^2} = 56.1$	$tan^{-1}\left(\frac{-21}{52}\right) = -22^0$
5	X-dir = 97-143 = -46 Y-dir = 93-141 = -48	$\sqrt{(-46)^2 + (-48)^2} = 66.5$	$tan^{-1}\left(\frac{-48}{-46}\right) = 46.2^0$
6	X-dir = 88-86 = 2 Y-dir = 87-87 = 0	$\sqrt{(2)^2 + (0)^2} = 1.4$	$tan^{-1}\left(\frac{0}{2}\right) = 0^0$
7	X-dir = 53-133 = -80 Y-dir = 112-56 = 56	$\sqrt{(-80)^2 + (56)^2} = 97.7$	$tan^{-1}\left(\frac{56}{-80}\right) = -35^0$
8	X-dir = 53-40 = 13 Y-dir = 62-52 = 10	$\sqrt{(13)^2 + (10)^2} = 16.4$	$tan^{-1}\left(\frac{10}{13}\right) = 37.6^0$
9	X-dir = 85-52 = 33 Y-dir = 117-40 = 77	$\sqrt{(33)^2 + (77)^2} = 83.8$	$tan^{-1}\left(\frac{77}{33}\right) = 66.8^0$
10	X-dir = 190-146 = 44 Y-dir = 185-120 = 65	$\sqrt{(44)^2 + (65)^2} = 78.5$	$tan^{-1}\left(\frac{65}{44}\right) = 55.9^0$
11	X-dir = 117-87 = 30 Y-dir = 85-73 = 12	$\sqrt{(30)^2 + (12)^2} = 32.3$	$tan^{-1}\left(\frac{12}{30}\right) = 21.8^0$
12	X-dir = 33-117 = -84 Y-dir = 85-68 = 17	$\sqrt{(-84)^2 + (17)^2} = 85.7$	$tan^{-1}\left(\frac{17}{-84}\right) = -11.4^0$
13	X-dir = 95-33 = 62 Y-dir = 45-106 = -61	$\sqrt{(62)^2 + (-61)^2} = 87.0$	$tan^{-1}\left(\frac{-61}{62}\right) = -44.5^0$
14	X-dir = 38-153 = -115 Y-dir = 29-177 = -148	$\sqrt{(-115)^2 + (-148)^2} = 187.4$	$tan^{-1}\left(\frac{-148}{-115}\right) = 52.2^0$
15	X-dir = 205-90 = 115 Y-dir = 142-171 = -28	$\sqrt{(115)^2 + (-28)^2} = 118.4$	$tan^{-1}\left(\frac{-28}{115}\right) = -13.7^0$
16	X-dir = 122-80 = 42 Y-dir = 80-75 = 5	$\sqrt{(42)^2 + (5)^2} = 42.3$	$tan^{-1}\left(\frac{5}{42}\right) = 6.8^0$
17	X-dir = 66-118 = -52 Y-dir = 87-111 -24	$\sqrt{(-52)^2 + (-24)^2} = 57.3$	$tan^{-1}\left(\frac{-24}{-52}\right) = 24.8^0$
18	X-dir = 61-71 = -10 Y-dir = 99-46 = 52	$\sqrt{(-10)^2 + (52)^2} = 53.9$	$tan^{-1}\left(\frac{52}{-10}\right) = -79.3^0$
19	X-dir = 118-63 = 55 Y-dir = 165-139 = 26	$\sqrt{(55)^2 + (26)^2} = 60.8$	$tan^{-1}\left(\frac{26}{55}\right) = 25.3^0$

20	X-dir = 192-139 = 53 Y-dir = 184-161 = 23	$\sqrt{(53)^2 + (23)^2} = 57.8$	$tan^{-1}\left(\frac{23}{53}\right) = 23.5^0$
21	X-dir = 76-74 = 2 Y-dir = 73-70 = 3	$\sqrt{(2)^2 + (3)^2} = 3.6$	$tan^{-1}\left(\frac{3}{2}\right) = 56.3^0$
22	X-dir = 55-86 = -31 Y-dir = 106-71= 35	$\sqrt{(-31)^2 + (35)^2} = 46.8$	$tan^{-1}\left(\frac{35}{-31}\right) = -48.5^0$
23	X-dir = 55-68 = -13 Y-dir = 65-47 = 18	$\sqrt{(-13)^2 + (18)^2} = 22.2$	$tan^{-1}\left(\frac{18}{-13}\right) = -78.7^0$
24	X-dir = 93-83 = 10 Y-dir = 134-36 = 36	$\sqrt{(10)^2 + (36)^2} = 37.4$	$tan^{-1}\left(\frac{10}{36}\right) = 74.5^0$
25	X-dir = 153-122 = 31 Y-dir = 151-139 = 12	$\sqrt{(31)^2 + (12)^2} = 33.2$	$tan^{-1}\left(\frac{12}{31}\right) = 21.2^0$

Source: Authors' computation

Table 3: Gradients' Splitting into Bins

Bins	Gradient Direction	Frequency	Gradient Magnitude using weighted voting
0	$-80^0 - -62^0$	3	$(-79.3 + -79.2 + -78.7) / 3 = -79.1$
1	$-62^0 - -44^0$	2	$(-48.5 + -44.5) / 2 = -46.5$
2	$-44^0 - -26^0$	1	$-35/1 = -35$
3	$-26^0 - -8^0$	4	$(-22 + -18.4 + -13.7 + -11.4) / 4 = -16.4$
4	$-8^0 - 10^0$	2	$(0 + 6.8) / 2 = 3.4$
5	$10^0 - 28^0$	5	$(21.2 + 21.8 + 23.5 + 24.8 + 25.3) / 5 = 23.3$
6	$28^0 - 46^0$	2	$(37.6 + 45) / 2 = 41.3$
7	$46^0 - 64^0$	4	$(46.2 + 52.2 + 55.9 + 56.3) / 4 = 52.7$
8	$64^0 - 82^0$	2	$(66.8 + 74.5) / 2 = 70.7$
	Total frequency	25	

Source: Jupyter Notebook Code Editor

Finally, the size of extracted features is calculated using the number of blocks in overlapping block processing This is achieved by:

$$Number\ of\ blocks\ = \frac{(Image\ size - Block\ size)}{Stride} + 1 \quad ...5$$

$$Number\ of\ blocks\ (rows) = \frac{(16-2)}{1} + 1 = 15$$

$$Number\ of\ blocks\ (columns) = \frac{(16-2)}{1} + 1 = 15$$

For normalization purpose, a 36-point feature vector is collected. Therefore,

$$size\ of\ features\ extracted = size\ of\ horizontal\ blocks * size\ of\ vertical\ bocs * feature\ vector\ points \quad ...6$$

$$= 15 * 15 * 36$$

$$= 8100$$

Table 4: Size of the Extracted Features

Extracted Features	Dimension	Size
Original Features	(128*128*3)	49152
Extracted Features	(15*15*36)	8100

Source: Authors' computation

There are 8100 extracted from the total 49152 features from Table 4.

4.3 Dataset experimentation

Table 5: Exploratory Data Analysis for the Experimental Analysis of Facial Features

	Pixel Value		**Proportion**	
Facial Attribute	**People with drug addiction**	**People without drug addiction**	**People with drug addiction**	**People without drug addiction**
Standard Deviation				
Left-eye	125.635186	39.468975	0.420214	0.242652
Right-eye	125.635186	39.468975	0.420214	0.242652
Mouth	254.810322	97.199794	0.661642	0.593692
Nose	149.209584	56.305417	0.308172	0.344848
Left-cheek	283.148018	64.805864	0.785226	4199.8
Right-cheek	283.148018	60.350642	0.785226	3642.2
Variance				
Left-eye	15784.20000	1557.8	0.17658	0.05888
Right-eye	15784.20000	1557.8	0.17658	0.05888
Mouth	64928.30000	9447.8	0.43777	0.35247
Nose	22263.50000	3170.3	0.09497	0.11892
Left-cheek	80172.80000	0.393599	0.61658	0.15492
Right-cheek	80172.80000	0.367859	0.61658	0.13532

Source: Authors' computation

Table 5 reveals that the standard deviations for people with drug addiction are significantly higher than those without drug addiction for all facial attributes:

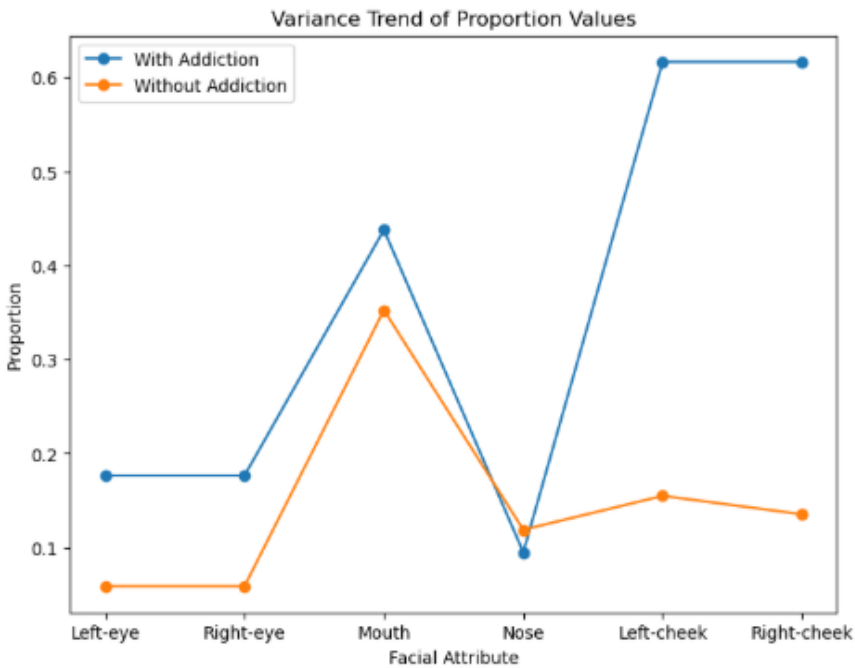
- Left-eye & Right-eye: 125.64 (with addiction) vs. 39.47 (without addiction)
- Mouth: 254.81 (with addiction) vs. 97.20 (without addiction)
- Nose: 149.21 (with addiction) vs. 56.31 (without addiction)
- Left-cheek & Right-cheek: 283.15 (with addiction) vs. 0.79 (without addiction)

Equally, the variance values for the "with addiction" group are significantly higher than those for the "without addiction" group for all facial attributes.

- Left-eye & Right-eye: 15784.2 (with addiction) vs. 1557.8 (without addiction)
- Mouth: 64928.3 (with addiction) vs. 9447.8 (without addiction)
- Nose: 22263.5 (with addiction) vs. 3170.3 (without addiction)
- Left-cheek: 80172.80000 (with addiction) vs. 0.393599 (without addiction)
- Right-cheek: 80172.80000 (with addiction) vs. 0.367859 (without addiction)

Figure 8 represents the significance difference between the variance trend of proportion values in people with and without drug addiction.

Figure 8: Standard Deviation Trend of Proportion Values



Source: Jupyter Notebook Code Editor

Inference: The rationale for the difference in variance between the "with addiction" and "without addiction" groups can be attributed to the following feature-specific justifications.

- Left and Right Eyes: The significantly higher variance in eye-related among individuals with drug addiction features might indicate that drug addiction could lead to changes in eye health, such as redness, puffiness, or other signs, contributing to increased variability.

- Mouth: The larger variance in mouth-related features among individuals with drug addiction could be due to drug addiction influence on facial expressions, leading to increased variability.
- Nose: The lower variance in nose-related features among individuals with drug addiction might suggest that drug addiction does not necessarily lead to increased variability in nasal health or appearance.
- Left and right Cheeks: The higher variance in cheek appearance among individuals with drug addiction, could be related to substance abuse effects.

4.4 Results from feature selection

Table 6 reveals that 50 features were selected out of the 8900 extracted features.

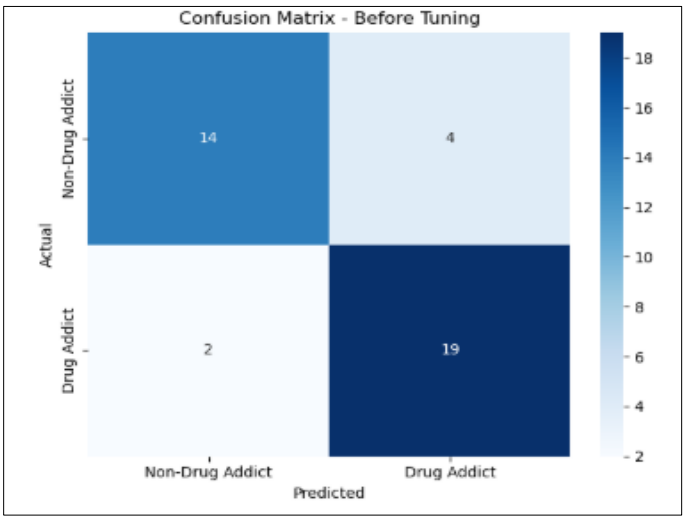
Table 6: Feature Selection

Features	Size
Extracted Features	8100
Selected Features	50

4.5 Results from model's performance evaluation

Figures 9 and 10 represent the confusion matrix before and after tuning, respectively.

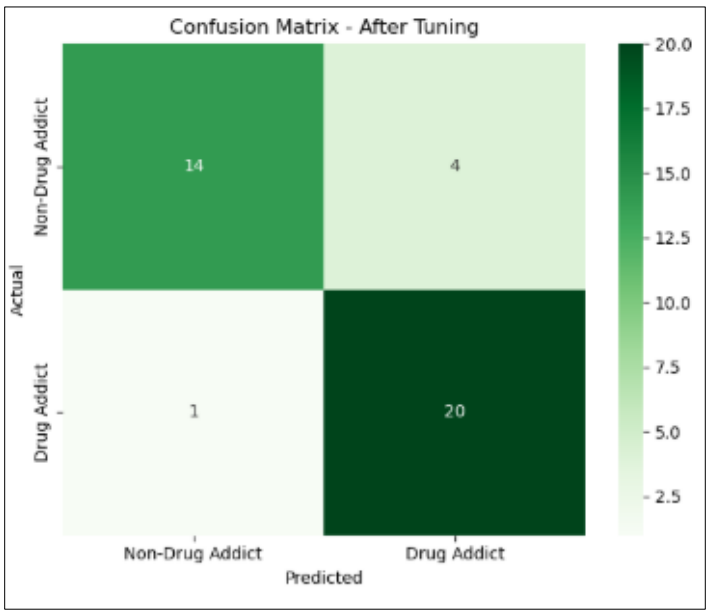
Figure 9: Confusion Matrix before Hyperparameter Tuning



Source: Jupyter Notebook Code Editor

From Figure 9, TP = 19, TN=14, FP =4, and FN= 2. From Figure 10, TP= 20, TN = 14, FP = 4, and FN = 1.

Figure 10: Confusion Matrix after Hyperparameter Tuning



Source: Jupyter Notebook Code Editor

Table 7: Summary of the Model's Evaluation

Metric	Before Hyperparameter Tuning	After Hyperparameter Tuning	Change
Values	TP = 19, TN = 14, FP = 4 and FN = 2	TP = 20, TN = 14, FP = 4 & FN = 1	
Accuracy	$\frac{19 + 14}{19 + 14 + 4 + 2} = 84.62\%$	$\frac{20 + 14}{20 + 14 + 4 + 1} = 87.18$	2.56% ↑
Precision	$\frac{19}{19 + 4} = 82.61\%$	$\frac{20}{20 + 4} = 83.33\%$	0.72% ↑
Recall	$\frac{19}{19 + 2} = 90.48\%$	$\frac{20}{20 + 1} = 95.24\%$	4.76% ↑
F1-score	$2 * \left(\frac{0.8261 * 0.9048}{0.8261 + 0.9048}\right) = 86.37\%$	$2 * \left(\frac{0.8333 * 0.9524}{0.8333 + 0.9525}\right) = 88.89\%$	2.52% ↑

Table 7 represents the RFC model's evaluation result. The accuracy of the model before hyperparameter tuning was 84.62%, which increased to 87.18% after tuning. This represents a 2.56% improvement, indicating that the model is better at correctly classifying instances. The precision of the model before hyperparameter tuning was 82.61%, which

slightly increased to 83.33% after tuning. This represents a 0.72% improvement, showing that the model is slightly more accurate in identifying true positives. The recall of the model before hyperparameter tuning was 90.48%, which significantly increased to 95.24% after tuning. This represents a 4.76% improvement, indicating that the model is much better at detecting actual positive instances. The F1 score of the model before hyperparameter tuning was 86.37%, which increased to 88.89% after tuning. This represents a 2.52% improvement, showing that the model's balance between precision and recall has improved. Overall, the hyperparameter tuning process has resulted in significant improvements in the model's performance across all metrics.

Table 8: Results Comparison with Existing Studies

Authors	Methodology	Accuracy	Precision	Recall	F1-score
Haque *et al.* (2021)	**Algorithm**: RF+Pearenson correlation **Dataset**: YMM dataset	95%	99%	44%	-
Choi *et al.* (2021)	**Algorithm**: RF+PCA+Chi-square **Dataset**: 2019 National Survey on Drug Use and Health survey	99.8%	99.2%	100%	100%
Choi *et al.* (2021)	**Algorithm**: RF + Relief **Dataset**: National Youth Tobacco Survey (2019) dataset	73.4%	-	-	-
Arif *et al.* (2021)	**Algorithm**: RF+PCA **Dataset**: Primary dataset	74%	52%	63%	81%
Current Study	**Algorithm**: RFC+GCSV **Dataset**: locally sourced dataset	≈ 87%	≈ 83%	≈ 95%	≈ 89%

Source: Authors' computation

Table 8 informs that integrating Random Forest (RF) with other techniques yields varying levels of performance. The combination of RF with Pearson correlation by Haque *et al.* (2021) reports high accuracy and precision but low recall. The integration of RF with PCA and Chi-square by Choi *et al.* (2021) shows exceptional performance with accuracy and precision above 99%. The combination of RF with Relief by Choi *et al.* (2021b) reports a moderate accuracy of 73.4%. this study's integration of RFC with GCSV shows promising results with an accuracy of approximately 87% and balanced precision (83%), recall (95%) and f1-score (89%).

5.0 Conclusion and Recommendations

This study developed and presented an improved facial analysis-based drug addiction prediction system using a machine learning algorithm by collecting and

preprocessing facial images, extracting and selecting relevant features, and developing and optimizing a model using the random forest classification algorithm and grid search cross-validation technique, respectively.

The results of this study demonstrate the effectiveness of the drug addiction prediction system in predicting drug addiction. Additionally, the analysis of facial features revealed significant differences between individuals with and without drug addiction. This study contributes to knowledge in the following ways:

- Utilization of a locally collected dataset, providing insights into drug addiction in a specific context.
- Employment of Histogram of Oriented Gradients (HOG) for feature extraction and Recursive Feature Elimination (RFE) for feature selection, enhancing model performance.
- Optimization of random forest classification algorithm parameters using grid search cross-validation, improving model accuracy.
- Exploration of differences in facial features between individuals with and without drug addiction, shedding light on potential biomarkers.

Hence, these contributions advance the understanding of facial analysis-based drug addiction prediction and provide a foundation for future research. To enhance the model's performance and generalizability, future research can source a huge dataset from healthcare institutions and employ deep learning technique to increase the model's accuracy and robustness.

References

Almahmood, M., Najadat, H., Alzu'bi, D., Abualigah, L., Zitar, R. A., Abualigah, S., & Al-Saqqar, F. (2023). Predictive model of psychoactive drugs consumption using classification machine learning algorithms. *Applied and Computational Engineering*, *8*(1), 738-743.

Amato, G., Falchi, F., Gennaro, C., & Vairo, C. (2018). A comparison of face verification with facial landmarks and deep features. In *International Conference on Advances in Multimedia (MMEDIA)* pp. 1–6.

Arif, M., Alghamdi, K. K., Sahel, S. A., Alosaimi, S. O., Alsahafi, M. E., Alharthi, M. A. & Arif, M. (2021). Role of machine learning algorithms in forest fire management: A literature review. *Journal of Robotics & Automation, 5*(1), 212-226.

Basuni, N. & Siregar, A. M. (2023). Comparison of the accuracy of drug user classification models using machine learning methods. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *7*(6), 1348-1353.

Chavan, M., Yawalkar, Y. & Suryawanshi, K. (2024). Innovation in healthcare by implementing IoT application to achieve sustainability: A study from Indian perspective. *Computology: Journal of Applied Computer Science and Intelligent Technologies, 4* (1), 55-73. Retrieved From: https://www.journalpressindia.com/computology-journal-of-applied-computer-science-and-intelligent-technologies/doi/10.17492/computology.v4i1.2404

Choi, J., Chung, J. & Choi, J. (2021). Exploring impact of Marijuana (Cannabis) abuse on adults using machine learning. *International Journal of Environmental Research and Public Health*, *18*(19), 1-12.

Choi, J., Jung, H. T., Ferrell, A., Woo, S. & Haddad, L. (2021b). Machine learning-based nicotine addiction prediction models for youth e-cigarette and waterpipe (Hookah) users. *Journal of Clinical Medicine*, *10*(5), 1-13.

Deedee, B., Onate, T., & Emmah, V. (2024). Fake profile detection and stalking prediction on X using random forest and deep convolutional neural networks. *Computology: Journal of Applied Computer Science and Intelligent Technologies, 4* (1), 1-19. Retrieved From: https://www.journalpressindia.com/computology-journal-of-applied-computer-science-and-intelligent-technologies/doi/10.17492/computology.v4i1.2401

Gong, H., Xie, C., Yu, C., Sun, N., Lu, H., & Xie, Y. (2021). Psychosocial factors predict the level of substance craving of people with drug addiction: A machine learning approach. *International Journal of Environmental Research and Public Health*, *18*(22), 1-12.

Gu, X., Yang, B., Gao, S., Yan, L. F., Xu, D., & Wang, W. (2021). Application of bi-modal signal in the classification and recognition of drug addiction degree based on machine learning. *Mathematical Biosciences and Engineering, 18*(5), 6926-6940.

Haque, U. M., Kabir, E. & Khanam, R. (2021). Detection of child depression using machine learning methods. *PLoS one*, *16*(12), 1-13.

Jeon, H. & Oh, S. (2020). Hybrid-recursive feature elimination for efficient feature selection. *Applied Sciences*, *10*(9), 1-8.

Lakshmi, J. V. N. & Das, A. (2023). Forecasting the risk of coronary heart diseases using machine learning algorithms. *Computology: Journal of Applied Computer Science and Intelligent Technologies, 3*(2), 133-152. Retrieved From https://www.journalpressindia.com/computology-journal-of-applied-computer-science-and-intelligent-technologies/doi/10.17492/computology.v3i2.2307

Li, Y., Yan, X., Zhang, B., Wang, Z., Su, H. & Jia, Z. (2021). A method for detecting and analyzing facial features of people with drug use disorders. *Diagnostics*, *11*(9), 1-18.

Mahmoud, B. S. & Garko, A. B. (2022). A machine learning model for malware detection using recursive feature elimination (RFE) for feature selection and ensemble technique. *IOS Journals*, *24*(1), 23-30.

Misra, P. & Yadav, A. S. (2020). Improving the classification accuracy using recursive feature elimination with cross-validation. *International Journal of Emerging Technologies*, *11*(3), 659-665.

Oliva, V., De Prisco, M., Pons-Cabrera, M. T., Guzmán, P., Anmella, G., Hidalgo-Mazzei, D., Grande, I., Fanelli, G., Fabbri, C., Serretti, A., Fornaro, M., Lasevoli, F., de Bartolomeis, A., Murru, A., Vieta, E. & Fico, G. (2022). Machine learning prediction of comorbid substance use disorders among people with bipolar disorder. *Journal of Clinical Medicine*, *11*(14), 1-13.

Parekh, T. & Fahim, F. (2021). Building risk prediction models for daily use of marijuana using machine learning techniques. *Drug and Alcohol Dependence*, *225*, 1-6.

Ramezan, C. A. (2022). Transferability of recursive feature elimination (RFE)-derived feature sets for support vector machine land cover classification. *Remote Sensing*, *14*(24), 1-25.

Sharma, A., Sharma, H., Varshney, S. & Gusain, N. (2023). Heart disease prediction: A comparative analysis of machine learning algorithms. *Computology: Journal of Applied Computer Science and Intelligent Technologies, 3*(2), 171-192. Retrieved From https://www.journalpressindia.com/computology-journal-of-applied-computer-science-and-intelligent-technologies/doi/10.17492/computology.v3i2.2309

Uddin, M. N., Hafiz, M. F. B., Hossain, S. & Islam, S. M. M. (2022). Drug sentiment analysis using machine learning classifiers. *International Journal of Advanced Computer Science and Applications*, *13*(1), 92-100.

UNODC (2022). World drug report 2018. Executive summary conclusions and policy implications. Retrieved from https://www.unodoc.org.wdr