Received: 14 June 2025 | Revised: 19 June 2025 | Accepted: 23 June 2025 | Vol. 1 | Issue 1 | June 2025
e-ISSN: 3105-2169

International Journal of Mechatronics, Robotics, and Artificial Intelligence
Original article

Utilizing Machine Learning Algorithms and SMOTE for Analyzing and Predicting Homicides

Hussain A. Younis [1,2 *], **Ghazwan abdulnabi** [3], **Israa M. Hayder** [4], **Sani Salisu** [5], **Maged Nasser** [6]
[1] College of Education for Women, University of Basrah, Basrah 61004, Iraq
[2] School of Computer Sciences, Universiti Sains Malaysia, Gelugor 11800, Malaysia
[3] Department of Computer Science (Educational Science), University of Basrah, 61004, Basrah, Iraq
[4] Department of Computer Systems Techniques, Qurna Technique Institute, Southern Technical University, Basrah, 61016, Iraq
[5] Department of Information Technology, Federal University Dutse, Dutse 720101, Nigeria
[6] Positive Computing Research Group, Institute of Autonomous Systems, Department of Computing, Universiti Teknologi PETRONAS, 32610 Seri Iskandar, Perak, Malaysia

Correspondence
*Hussain A. Younis
College of Education for Women, University of Basrah, Basrah 61004, Iraq
Email: hussain.younis@uobasrah.edu.iq

Abstract
This study analyzes homicide data in the United States from 1980 to 2014 using machine learning techniques to predict crime resolution and classify victim gender. The dataset, obtained from the FBI Supplementary Homicide Report, contains 638,454 records. Data preprocessing involved cleaning, converting categorical features to numerical values, and addressing class imbalance using Synthetic Minority Oversampling Technique (SMOTE). Various classification algorithms were applied, including Decision Tree and Naïve Bayes. The results showed that the Decision Tree model achieved 95% accuracy in predicting crime resolution and 85% accuracy in classifying victim gender, while Naïve Bayes reached 92% accuracy in crime resolution prediction. The findings highlight the effectiveness of machine learning in crime pattern analysis and prediction, aiding law enforcement in making more informed investigative decisions.

Keywords
Machine learning, Homicide analysis, SMOTE, Decision tree, Naïve Bayes, Classification algorithms.

I. Introduction

Crimes are all acts outside of morality and laws, which are deviant acts that deserve punishment and responsibility. And there is an assault on individuals, property, and society as a whole. The crimes vary in their forms and forms, from murder, robbery, rape, looting, pillage, banditry, and bribery. The law divided each crime with a penalty commensurate with the act committed by criminal crimes. Crimes were divided into categories based on several factors relating to the perpetrator of the act, the nature of the crime, the offender's circumstances, and the motives that led him. Crime figures indicate a secret spike in the rape rate of 22.6 percent, and 273 murders were registered in New York City as of December 2018 (NYC) [1].

They are not limited solely to murder. Each crime, however is based on a tangible physical structure without which it is not possible to achieve its legal structure. In all crimes, it seems that this structure does not take a single form, because the nature of the component acts depends on the type of crime, on the requirements of criminalisation. Crime, on the other hand, is natural and unpredictable, or it happens concurrently and in the same place and is not random. And after increasing the volume of criminal record data that can be used to analyse and predict current and future crimes, modern technologies such as deep learning and Geographical Information Systems (GIS), which are important for identifying hot spots and understanding the reasons behind the occurrence of these crimes, will make this possible [2] - [4]. Officials and decision-makers are working to predict, limit and prevent crime at these locations, which reduces social harm and stabilises state security. Moreover, the prediction of crime is better than the investigation into the course of that crime, which has varied in our modern era. There are many types of crimes, including thefts of all kinds, forgery, impersonation and electronic crimes, which are considered to be the most serious crimes of our modern age

[5]. This classification is binary, that is, the prediction of whether the problem will be solved. Moreover, victim gender will be classified. Because there are some victims whose gender is unknown, and this helps law enforcement officials know the gender of the victim. Prediction in comparative studies [6]-[8].

II. Dataset

This dataset was obtained as part of the United States' Justice Project initiative. It is a non-profit organization founded in 2015 that aims to collect and distribute information about unsolved murders from federal, state and local governments. It contains murder cases from the 1980-2014 FBI Supplementary Homicide Report. There was one dataset with all examples (638454 rows) and 24 inputs ordered by date from which database.csv is used (from January 1980 to September 2014) There are 24 variables for input and 1 variable for output (desired target). There were various types of crime details in the dataset such Record ID, City, State, Year, Month, Incident, Victim Age, Victim Race, etc. and one output variable y denotes whether the crime-solving or not. Such datasets were loaded in python language for Pre-processing to check for any missing values and found there are unknown values. However, the dataset is imbalanced. Homicide Incidents, 1980-2014 The dataset was collected from Source: www.kaggle.com/murderaccountability/homicide-reports [9]-[11].

A. Pre-processing

In this step, we need to process our data. This is done in four steps: the first step, the drop of unknown data, the second step is to convert features from Categorical to Numeric, the third step is to balance the data, and the last step is Dimensional reduction.

B. Clean dataset

In the beginning, we must clean the data from the null values. In our database, because some attributes contain unknown values. In Table I, the data set contains organized data connected to delinquent events, to administrational, demographic and relational arenas. It has such fields as Record ID, Agency Code, Agency Name and Agency Type that allude to the reporting agency and its location, and to such time indicators as Year and Month. Specific crime information is recorded in such columns as Crime Type, Crime Solved, Weapon. Victim-related information contains Victim Sex; Victim Age; Victim Race; Victim Ethnicity, and Victim Count. Equally, data on victimization is documented in Perpetrator Sex, Perpetrator Age, Perpetrator Race, Perpetrator Ethnicity and Perpetrator Count. The column heading Relationship explains the nature of the relationship between the victim and the perpetrator and the column name Record Source explains on which basis, the record was started. The columns are Integer format (int64). The missing values are implemented differently with a greater amount in some fields such as the field of ethnicity and relationship, therefore, it will be necessary to account them during analysis. You must dispose of these values because they affect the classifier's performance and are considered noise values.

TABLE I.
THE NULL VALUES

Record ID	0
Agency Code	0
Agency Name	47
Agency Type	0
City	0
State	0
Year	0
Month	0
Incident	0
Crime Type	0
Crime Solved	0
Victim Sex	984
Victim Age	0
Victim Race	6676
Victim Ethnicity	368303
Perpetrator Sex	190365
Perpetrator Age	1
Perpetrator Race	196047
Perpetrator Ethnicity	446410
Relationship	273013
Weapon	33192
Victim Count	0
Perpetrator Count	0
Record Source	0
dtype: int64	

As it turns out, there are a lot of empty values. If the record is deleted, this leads to the worst prediction. If we suggest deleting the feature entirely, then some of the features are half the features empty. Therefore, 'Victim Ethnicity', 'Perpetrator Ethnicity', 'Perpetrator Sex', 'Perpetrator Race', 'Weapon', 'Perpetrator Count', 'Record Source. And then Record ID', 'Agency Code', ' Agency Name ',' Agency Type 'because these are random values. The features obtained are shown in Table II.

TABLE II.
FEATURES OBTAINED

City	0
State	0
Year	0
Month	0
Incident	0
Crime Type	0
Crime Solved	0
Victim Sex	984
Victim Age	0
Victim Race	6676
Perpetrator Age	1
Relationship	273013
Victim Count	0

C. *Transformation the dataset*

As we mentioned earlier, there are some categorical data that we need to convert to numeric as in Table III. Contrast two forms of the same data set; unprocessed and processed. The original data (left side) consists of different types of data with the categorical (the fields storing City, State, Crime Type are stored as object) and numerical (fields of these types are stored as int64 or float64) parameters. It also contains some columns of missing data like Victim Sex, Victim Race, and Relationship, this property is not ideal in terms of usage in machine learning models. On the contrary, the preprocessed dataset (right side) has a clean and consistent structure. The categorical columns will also be cast to int32 as the numeric columns, and this will mean that some encoding methods (e.g. label encoding or one-hot encoding) are being used. Additionally, the missing records are dealt with and cannot be seen in non-constant non-zero counts on the columns. This conversion makes the dataset model-ready, numerically consistent and lacks missing data, which makes it suitable to train and test algorithms.

TABLE III.
DATASET TRANSFORMATION

#	Column	Non-Null Count	Dtype
0	City	638454 non-null	object
1	State	638454 non-null	object
2	Year	638454 non-null	int64
3	Month	638454 non-null	object
4	Incident	638454 non-null	int64
5	Crime Type	638454 non-null	object
6	Crime Solved	638454 non-null	object
7	Victim Sex	637470 non-null	object
8	Victim Age	637480 non-null	float64
9	Victim Race	631778 non-null	object
10	Perpetrator Age	638453 non-null	float64
11	Relationship	365441 non-null	object
12	Victim Count	638454 non-null	int64

#	Column	Non-Null Count	Dtype
0	City	637469 non-null	int32
1	State	637469 non-null	int32
2	Year	637469 non-null	int64
3	Month	637469 non-null	int32
4	Incident	637469 non-null	int64
5	Crime Type	637469 non-null	int32
6	Crime Solved	637469 non-null	int32
7	Victim Sex	637469 non-null	int32
8	Victim Age	637469 non-null	int64
9	Victim Race	637469 non-null	int32
10	Perpetrator Age	637469 non-null	int64
11	Relationship	637469 non-null	int32
12	Victim Count	637469 non-null	int64

1) *Dataset unbalanced*

When employing classification algorithms such as Decision Tree, Naïve Bayes, and similar models, their performance may degrade or result in overfitting, particularly when working with imbalanced datasets. If one class constitutes 70% of the data and the other only 30%, the accuracy metric becomes unreliable and may not reflect true model performance. Classifiers tend to favor the majority class, leading to biased predictions and reduced generalization. Consequently, the accuracy obtained from such models may be misleading and significantly lower than expected. Hence, it is essential to preprocess the dataset before applying classification algorithms. One effective technique SMOTE, which artificially generates new instances for the minority class to balance the dataset. This approach enhances the classifier's ability to learn from all classes more equally, as illustrated in Fig. 1.

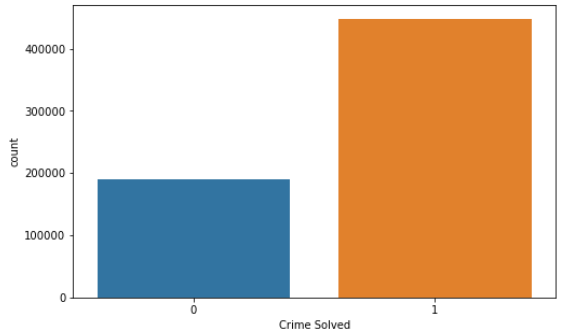

Fig. 1. Unbalanced dataset

2) *Synthetic SMOTE*

It is working on choosing a point from minority and then creating a new point by k-nearest neighbours good approach but requires a lot of calculations [6], [12]-[14] as shown in Fig. 2.

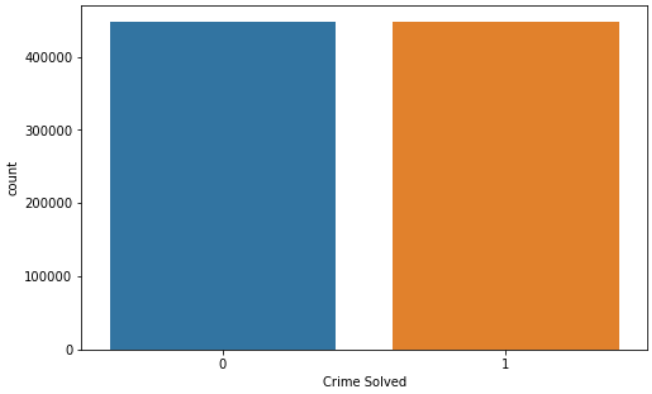

Fig. 2. Balanced dataset

III. MACHINE LEARNING ALGORITHMS

Two models will be applied to our dataset, which are Decision Tree and Nave Bayes as shown Fig. 3.

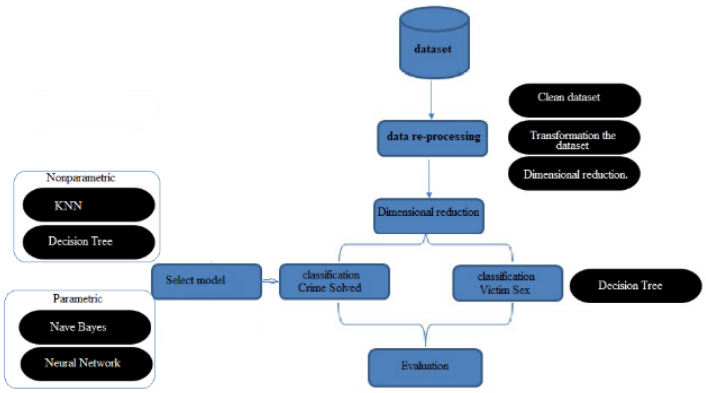

Fig. 3. Work methodology

In this step, the techniques used in this project applied in the Python environment will be presented. There are two objectives: The first objective is to classify whether they will be resolved crime or not, and the second is the gender of the victim.

A. Classifiers

After applying data processing in the previous steps, in this step, we will implement machine learning algorithms, which are divided into two: parametric and non-parametric classifiers.

1) Non-parametric classifier

It is also called lazy teaching, as it does not use assumptions in learning. Simply, the samples collected in training data are used. The algorithms that under this technique are Knn, decision trees, etc. After processing it in the previous steps, we will apply our dataset to the and decision trees.

2) Parametric classifier

Parametric algorithms are called linear machine learning algorithms, and linear regression is often used in them [15]. The algorithms that under this technique are Naive Bayes etc. After processing it in the previous steps, we will apply our dataset to the Naive Bayes classifiers.

B. Test Evaluation

Sometimes the accuracy scale does not give us an accurate description of the unbalanced data, so we need metrics to better insight the behaviour of the data.

A confusion matrix shown in Table IV is the result of a performance measure of a classification model that is used to summarize the outputs of the model in terms of four-way confusing results: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN), which can be used to analyze the key measures such as an accuracy, a precision, a recall, and F1-score to identify the effectiveness of a model and the nature of its mistakes.

TABLE IV.
CONFUSION MATRIX.

	Positive	Negative
Positive	TP	FP
Negative	FN	TN

Precision: the number of true positives divided by all positive predictions [6][16]-[20].

$$\text{Precision} = \frac{TP}{(TP+FP)} \qquad (1)$$

Recall: score combines precision and recall into one measure.

$$\text{TP rate} = \frac{TP}{P} \qquad (2)$$

F1: Score: measured Average of precision and recall.

$$Rate = \frac{FP}{N} \qquad (3)$$

C. Experiment Test Evaluation

After balancing the data, we have 895374 rows, and we will experiment with separating the data .30% for the testing dataset and 70% for the training dataset and random_state=5; as shown Fig. 4.

```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(data_pca, y, test_size=0.3, random_state=5, shuffle=False)
```

Fig. 4. Train_test_split (30%) and training (70%)

IV. EXPERIMENTS AND ANALYSIS

A. Experiments of Crime-Solved

The experiment was performed with Python, which contains most of the machine learning algorithms. Before modeling, we need a preliminary exploration of the data set (638454 rows and 24 features). We have previously explained that the data is not balanced to a large extent, and this problem was addressed using the (SMOTE) method, which was superior to the other methods. In this step, we will build the model (decision trees) for the non-parametric method, and we will try (Nave Bayes) for the parametric. And we will use the k-fold (5 folds) cross-validation that gave better results.

B. Decision Tree Model

Implemented the decision tree algorithm and obtained the highest accuracy after changing the model's settings, where the best criterion was chosen, which was entropy, and min_samples_split was 4. The length of the tree was 9 with k-fold (5 folds) cross-validation as shown Fig. 5.

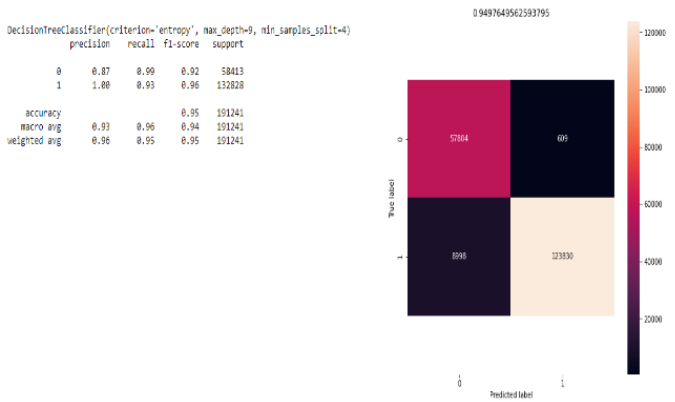

Fig. 5. Decision Tree model and classification report

C. Nave Bayes Model

Evaluation results given in Fig. 6 are of a statistical Gaussian Naive Bayes classifier based on binary classification of a dataset comprising 191,241 samples classified as either Class 0 samples (58,413 samples) or Class 1 samples (132,828 samples). Overall accuracy was 92.2% and performance metrics provided more details, having the Class 0 results based on precision and recall being 81%, and 98% respectively, and Class 1 results reaching almost perfect precision of 99%, with recall performance of 90%. The macro avg and the weighted avg F1-score attained the same value, of 0.91 and 0.92 respectively, which demonstrates good performance in both classes. The bottom confusion matrix is unmapped, so probably it denotes true/false prediction and Class 1 is better than Class 0. These outcomes show a good performance of the model, especially since the size of the dataset is large, and the classes are not balanced. The main advantages are high recall of Class 0 and excellent precision of Class 1, but the imbalance in precision-recall in Class 0 could be a matter of investigation based on the ability to accommodate the false positives expense requirement of the application. The scale-based metrics also verify the reliability of the model even in an unbalanced class.

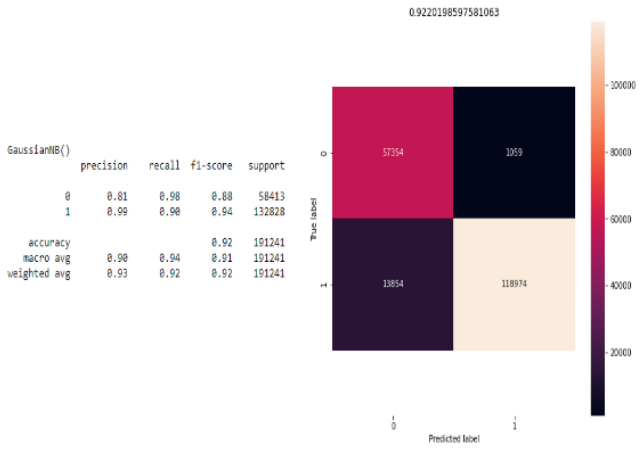

Fig. 6. Nave Bayes model and classification report.

D. Experiments of Victim-Sex

The benefit of this experience is to identify the gender of the victim in cases of severe thermal damage or advanced decomposition. In addition, the main goal is to have a group of crimes of unknown sex of the victim, which number 984 victims.

E. Decision Tree Model

Implemented the decision tree algorithm and obtained the highest accuracy after changing the model's settings, where the best criterion was chosen, which was Gini, and min_samples_splitwas 4. The length of the tree was 9 with k-fold (5 folds) cross-validation.

In Fig. 7, the results of the evaluation of a Decision Tree Classifier (max_depth=9, min_samples_split=4) on 191,241 instances, an example of which includes 41,216 case of Class 0, and of Class 1, due to 150,025 instances. The model reached a total accuracy score of 87.04%, but there was a significant difference in the results between the classes: where the model performed well with Class 1 (87% precision, 99% recall), it evidenced a comparatively lower result with Class 0 (91% precision and only 44% recall). The confusion matrix reports that 18,095 Class 0 and 168,267 Class 1 have been correctly assigned, 29,021 Class 0 and 1,758 Class 1 being misclassified. The weak performance of Class 0 influences macro average of 0.85, whereas the weighted average yields to Class 1, which is the largest. Such findings show a bias of the model in dealing with the majority class, which indicates the necessity of data imbalance reduction and parametrization to increase the general performance. The huge gap between precision and recall in Class 0 implies that there is a specific difficulty in accurately listing all incidents of this minority class.

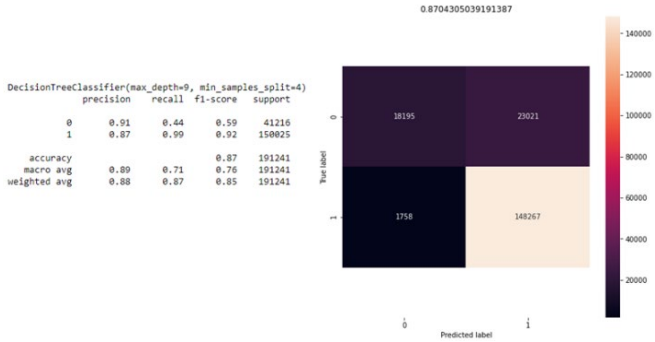

Fig. 7. Decision Tree model and classification report.

V. Results

Table V presents the performance outcomes of various models, along with the number of features utilized in each. The table includes evaluation metrics for two primary tasks: forecasting crime resolution and classifying victim gender. Among the tested models, the Decision Tree demonstrated the most favorable results. In predicting crime resolution, it achieved a testing accuracy of 95% and a pre-testing (anticipatory) accuracy of 96% using a maximum depth of 9. The Decision Tree model, characterized by its nonparametric structure, proved to be highly effective for this classification task.

In contrast, the Naïve Bayes model reached a testing accuracy of 92% and a pre-testing accuracy of 95% under default settings, indicating strong but slightly lower performance compared to the Decision Tree. For the victim gender classification task, the Decision Tree again outperformed Naïve Bayes. It recorded an accuracy of 85% during testing and 87% prior to testing, also with a maximum depth of 9. These findings indicate that the Decision Tree consistently outperformed Naïve Bayes in both crime resolution prediction and victim gender classification tasks. Overall, the Decision Tree model demonstrated greater robustness and predictive power, making it a more reliable choice for criminal data analysis.

TABLE V.
RESULTS OF DECISION TREE AND NAVE BAYES MODELS

problem	Model	Features selection	Value of parameter	Testing Accuracy	Score before testing	Type of model
Crime Solved	Decision tree	13	Depth = 9	95%	96%	n
	Nave Bayes	13	Default	92%	95%	p
Victim Sex	Decision tree		Depth =9	85%	87%	n

n: nonparametric, p: parametric

VI. CONCLUSION

The number of features used in the current assignment is 13 features instead of the actual 24 features because we analyse the features, we found that most of them are empty or random value that does not help in improving prediction. And then the reduced dimension from 13 D to 5 D by use reduced dimension Technique that improved the model's performance in terms of time speed. According to the task assigned to us, we had to choose only two algorithms, one parametric and the other referred to as non-parametric. The Naive Bayes algorithm was selected as a parametric algorithm and got an accuracy of 92%. Moreover, decision tree accuracy was (94%), considered non-parametric algorithms. Both models were used to predict crime solve whether or not, moreover, the victim sex was predicted using the decision tree, and we got 87% accuracy.

ACKNOWLEDGMENTS

We would like to thank the University of Basrah, Southern Technical University, Federal University Dutse, Universiti Teknologi PETRONAS and Universiti Sains Malaysia, incredibly if their assistance and cooperation. Their support and cooperation played a great role in the development and completion of this research.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest to report regarding the present study.

REFERENCES

[1] E. Lavanya, V. Mandalapu, and N. Roy. "Developing machine learning based predictive models for smart policing," *In 2019 IEEE International Conference on Smart Computing (SMARTCOMP)*, pp. 198-204, 2019, https://doi.org/10.1109/SMARTCOMP.2019.00053

[2] B. Gamze, and H. E. Colak. "Predicting and analyzing crime—Environmental design relationship via GIS - based machine learning approach," *Transactions in GIS* 28, no. 5, pp. 1377-1399, 2024, https://doi.org/10.1111/tgis.13195

[3] E.S. Mahimkar "Predicting crime locations using big data analytics and Map-Reduce techniques," *The International Journal of Engineering Research* 8, no. 4, pp. 11-21, 2021

[4] D. Fatima, M. Abu Talib, O. Abu Waraga, A. Bou Nassif, S. Abbas, and Q. Nasir, "Artificial intelligence & crime prediction: A systematic literature review," *Social Sciences & Humanities Open* 6, no. 1, 2022, https://doi.org/10.1016/j.ssaho.2022.100342

[5] B. Shruti, and R. K. Singh, "Exploration of Crime Detection Using Deep Learning," In *Innovations in Cyber Physical Systems: Select Proceedings of ICICPS 2020*, pp. 297-304. Springer Singapore, 2021, https://doi.org/10.1007/978-981-16-4149-7_26

[6] I. M. Hayder, G. A. Al Ali, and H. A. Younis, "Predicting reaction based on customer's transaction using machine learning approaches," *International Journal of Electrical and Computer Engineering*, vol.13, no. 1, 2023, https://doi.org/10.11591/ijece.v13i1.pp1086-1096

[7] N. A Sharma, A. S. Ali, and A. Kabir "A review of sentiment analysis: tasks, applications, and deep learning techniques," *International journal of data science and analytics*, vol. 19, no. 3, pp. 1-38, 2024, doi: 10.1007/s41060-024-00594-x

[8] H. A. Younis, N. I. R. Ruhaiyem, A. A. Badr, A. K. Abdul-Hassan, I. M. Alfadli, W. M. Binjumah, M. Nasser, "Multimodal age and gender estimation for adaptive human-robot interaction: A systematic literature review," *Processes*, vol. 11, issue 5, 2023, https://doi.org/10.3390/pr11051488

[9] S. Redkar, S. Mondal, A. Joseph, and K. S. Hareesha, "A Machine Learning Approach for Drug-target Interaction Prediction using Wrapper Feature Selection and Class Balancing," Mol. Inform., vol. 39, no. 5, 2020, doi: 10.1002/minf.201900062

[10] J. Ribeiro, L. Meneses, D. Costa, W. Miranda, and R. Alves, "Prediction of Homicide Urban Centers: A Machine Learning Approach," *under exclusive license to Springer Nature Switzerland AG 2022 K. Arai (Ed.): Intelligent Sys* 2021, LNNS, vol. 296, pp. 344–361, 2022. https://doi.org/10.1007/978-3-030-82199-9_22

[11] A. Shermila, A. B. Bellarmine, and N. Santiago, "Crime data analysis and prediction of perpetrator identity using machine learning approach," In *2018 2nd international conference on trends in electronics and informatics (ICOEI)*, pp. 107-114. IEEE, 2018, https://doi.org/10.1109/ICOEI.2018.8553904

[12] R. Geetha, S. Sivasubramanian, M. Kaliappan, S. Vimal, and S. Annamalai, "Cervical Cancer Identification with Synthetic Minority Oversampling Technique and PCA Analysis using Random Forest Classifier," *J. Med. Syst.,* vol. 43, no. 9, 2019, doi: 10.1007/s10916-019-1402-6

[13] J.H. Joloudari, A. Marefat, M.A. Nematollahi, S.S. Oyelere, & S. Hussain, "Effective class-imbalance learning based on SMOTE and convolutional neural networks," *Applied Sciences,* vol.13, issue 6, 2023, https://doi.org/10.3390/app13064006

[14] A. Fernández, S. Garcia, F. Herrera, & N.V. Chawla,"SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," Journal *of artificial intelligence research*, vol. 61, pp.863-905, 2023, https://doi.org/10.1613/jair.1.11192

[15] A. Géron, "*Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems," 2nd ed. Sebastopol, CA*, USA: O'Reilly Media, 2019.

[16] I. M. Hayder, T. A. Al-Amiedy, W. G. Saeed, M. Nasser, G. A. Al-Ali, and H. A. Younis. "An intelligent early flood forecasting and prediction leveraging machine and deep learning algorithms with advanced alert system," *Processes*, vol. 11, no. 2, 2023, https://doi.org/10.3390/pr11020481

[17] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, and M. Lang, "Benchmark for filter methods for feature selection in high-dimensional classification data," *Comput. Stat. Data Anal*., vol. 143, p. 106839, 2020, doi: 10.1016/j.csda.2019.106839

[18] A. Gupta,"A novel approach for classification of mental tasks using multiview ensemble learning (MEL)," *Neurocomputing*, vol. 417, pp. 558–584, 2020, doi: 10.1016/j.neucom.2020.07.050.

[19] H. A. Younis, N. I. Ruhaiyem, A. A. Badr, T. A. Eisa, M. Nasser, T. Tan , N.H. Samsudin, S. Salisu, "Creating the Hu-Int dataset: A comprehensive Arabic speech dataset for gender detection and age estimation of Arab celebrities," *Biomedical Signal Processing and Control*, vol.96, 2024, https://doi.org/10.1016/j.bspc.2024.106511

[20] J.T. Townsend, "Theoretical analysis of an alphabetic confusion matrix," *Perception & Psychophysics,* vol. 9, pp.40-50, 1971, https://doi.org/10.3758/BF03213026