2020 23rd International Conference on Computer and Information Technology (ICCIT)
19-21 December, 2020

# Intelligent Crime Investigation Assistance Using Machine Learning Classifiers on Crime and Victim Information

Saqueeb Abdullah[1], Farah Idid Nibir[2], Suraiya Salam[3], Akash Dey[4], Md Ashraful Alam[5] and Md Tanzim Reza[6]
[1,2,3,4,5,6]Department of Computer Science and Engineering, BRAC University, 66 Mohakhali, Dhaka 1212, Bangladesh
Email: [1]saqueeb10@gmail.com, [2]aidid.nibir97@gmail.com, [3]suraiyasalam.shitul@gmail.com, [4]ashomapto.akash@gmail.com, [5]ashraful.alam@bracu.ac.bd, [6]rezatanzim@gmail.com

2020 23rd International Conference on Computer and Information Technology (ICCIT) | 978-1-6654-2244-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/ICCIT51783.2020.9392668

*Abstract*—**In order to establish peace and justice in a society, it is essential to make proper and correct investigation of crime incidents. With the expansion of the utilization of computerized system to track crime and violence, computer applications can help law enforcement officers in a significant way. In most cases, crime incidents are kept in police database and these can be used for various helpful purpose. In this experiment, we have collected data of crime scenario from Bangladesh Police that had features such as area of crime, type of crime, number of victims and so on. Then we applied machine learning algorithms on the dataset for prediction of some attributes such as criminal age, sex, race, crime method etc. We used four different algorithms for our research: K-Nearest Neighbor (KNN), Logistic Regression (LR), Random Forest Classifier (RFC), Decision Tree Classifier (DTC). Using the aforementioned algorithms with 10 fold cross validation, we achieved different accuracy from all four attribute labels ranging from an average of approximate 75% to an average of approximate 90%. Despite the clear need of further improvement, the results give clear implications that it is possible to achieve well performing automated system for suspect attribute prediction with further work. Finally, we ended the research by comparing and analyzing all the achieved results.**

*Index Terms*—**Crime, investigation, Automated system, Classification, Features, Labels**

## I. Introduction

Criminal investigation is a multifaceted problem solving challenge. During investigation, an expert official is often required to examine the location of the crime. The official meticulously examines various important aspects of the crime scene, collects data and eventually analyzes data in order to infer identification information of the criminal. This complicated process of criminal identification demands high critical and reasoning skills. Additionally, most of the time these procedures are needed to be performed fairly quickly since criminals always try to hide all their traces. Therefore, the more time criminals get, the harder it becomes to track him down. In order to address all these complications, the crime scene examiners need to earn lots of experience and analytical skills so that they can make proper use of insightful information. [1] However, very few can earn such interpretative skills which results in a low number of proficient criminal investigators. Therefore, a lack of enough crime investigator is often evident. This is especially true for a country like Bangladesh, where the amount of crime is regularly growing and is expected to grow in the future along with the continuous growth of population. [2]

In this age of vast digitalization, various machine based approaches are being taken to automate the problem solving procedures across different fields. This automation of problem solving requires some typical steps such as collecting raw data, denoising it, analyzing it through computing machines and so on. These problem solving procedures are often referred as automated data driven approach where the data are being analyzed by a machine instead of a human. Criminal investigation methodologies are also mostly data driven as various data from the crime scene are used to deduct criminal information. Consequently, it is possible to apply machine based approach in this investigation.

In this experiment, we applied few machine learning algorithms to determine criminal attributes from crime information and compared between the results of the algorithms. The paper is divided into five main sections. In the next section, literature review of the proposed methodology is discussed. The following parts consist of the dataset details, proposed model, result and analysis, conclusion and discussion.

## II. Literature Review

### A. Previous Works

There is quite a few work that has been done for automated crime investigation in the past. Among the few ones that have been done, most of them used some form of data mining technology. Some of these data mining works include, usage of semi-supervised machine learning algorithms such as K-means clustering in order to discover essential knowledge from records of crime [3], usage of different types of regression algorithms to predict violent crime patterns from data [4], usage of data mining for fraud detection [5], prediction of event outcome through analyzing a dataset of criminal activity [6] and so on. Additionally, work has been done on predicting crime based on geographical features [7], urban planning features [8] etc. All these algorithms predict criminal attributes from a set of specific information that are often difficult to collect. From the perspective of Bangladesh, most of the research are crime forecast based. As a result, this type of criminal attribute predicting research has not been done before as there is no regulation of collecting and storing crime

978-1-6654-2244-4/20/$31.00 ©2020 IEEE

data. Therefore, there is a dire need of more crime prediction research in Bangladesh.

### B. Algorithms

The four different algorithms we have used for our research are classification algorithms that try to classify labels based on a feature set. LR is a regression technique that converts output to binary by using sigmoid function. When LR is used to classify between more than two classes, it is called as multinomial LR. [9] In our research, we have used multinomial LR for every label classification except gender, as gender has two value types: male and female. KNN on the other hand, is a clustering algorithm that tries to group together similarly labeled data into the same cluster. The value of K in KNN determines the number of nearest data point it tries to cluster together. [10] Furthermore, DTC is a classifier that creates tree structured branching shape based on different attributes for classification. [11] For our experiment, we used CART decision tree. [12] CART uses a metrics called gini index for classification. Finally, RFC algorithm creates a group of small classification trees with different branching attributes and combines them for very strong predictive power. [13]

## III. Proposed model

The proposed model starts with collection of the database and afterwards, some of the pre-processing steps were performed on the dataset. Then the dataset was divided into feature and label set. A portion of the feature set was used to train the machine learning classifiers and those classifiers tried to predict the labels. Before training, the entire dataset was divided into 80% train data and 20% test data. Subsequently, the feature set of the data was scaled and passed through four different classification algorithm: KNN, LR, RFC and DTC. Finally, all the different results were compared and analyzed.

## IV. Dataset details and processing

### A. Dataset details

We collected a completely new dataset for our research. The data were directly collected from Bangladesh Police under the Ministry of Home Affairs of the Government of Bangladesh. This dataset is difficult to find since it is classified data and full of critical information. Although the amount of samples in the dataset was not huge, there was still modest amount of sample just good enough to serve our purpose.

There were five different types of features in the dataset and there were four different types of corresponding labels alongside with it.

### B. Dataset processing

The raw dataset had some defects in it so those had to be resolved through some pre-processing steps. First of all, the rows with at least one empty value had to be taken care of. As criminal prediction is a critical task, we decided to drop entire rows that contained one or more null values. Afterwards, as labels such as 'age' had lots of different numerical values, the amount of variance was reduced by putting them into specific ranges. Finally, all the data points were encoded into numerical form from their string from for proper classification purpose. The details of all the features and labels of the dataset is given in table number I.

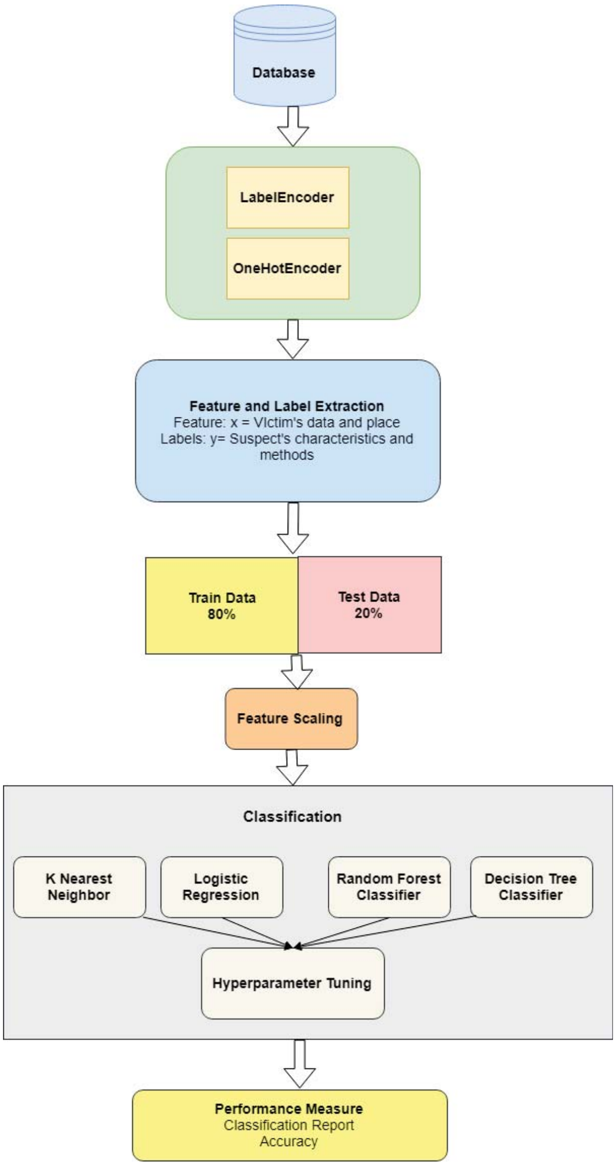

Fig. 1. Proposed model

As it is visible from table I, there were three different types of features in the dataset and four different types of labels. There were exactly 1466 data samples after the pre-processing steps were done. The data samples were divided into approximately 80% training data and 20% testing data for supervised learning purpose. As a result, 1172 data samples went into the training set and rest went into the testing set. During learning process, we took all the five features and one of the four labels at a time for classification purpose. During training period, we also applied exhaustive gridsearch on the parameters to find the the best parameters that can provide the most accurate results for each class. Additionally, we have

TABLE I
DETAILS OF FEATURES AND LABELS IN THE DATASET

| Type | Name | Types of values |
|---|---|---|
| Features | Area of crime | Sutrapur, Gulshan, Lalbagh, Adabor, Rampura, Mirpur, Shahbag, Bangsal, Hazaribagh, Motijheel and others |
| | Type of crime | Kidnap, Rape, Aggravated assault, Arson, Drug trafficking, False pretences, Embezzlement, Robbery, Terrorism, Murder and others |
| | Victim Sex | Male and Female |
| | Victim Race | White, Black and Brown |
| | Number of victims | 1-7 |
| Labels | Criminal Age | 31-40, 41-50, 51-60 and others |
| | Criminal Sex | Male and Female |
| | Criminal Race | White, Black and Brown |
| | Methods of the crimes | Firing, Unknown, Deadly weapon, Explosion, Bombing, Chloroform and others |

performed cross validation during training in order to avoid baised split of train-test dataset. We intentionally used the same set of train and test data for each of the classification algorithm during cross validation so that the results can be compared properly and accurately.

## V. RESULT AND ANALYSIS

After performing 10 fold cross validation on the dataset, we extracted some results for all the four predictive attributes. The accuracy measurements for methods of crimes are as follows,

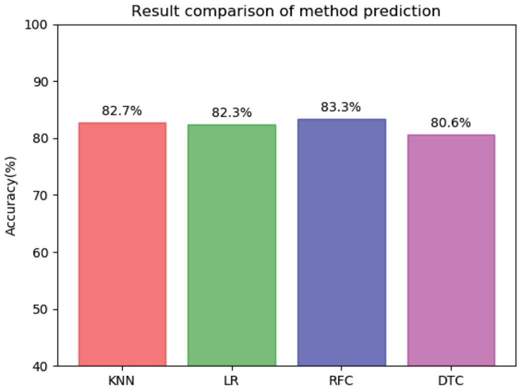

Fig. 2. Comparison between accuracy of four models (1)

As we can see from the figure 2, RFC achieves the best classification accuracy in case of method prediction. On the other hand, we found the lowest result from DTC. However, the results were mostly close to each other.

On other hand, in figure 3, again RFC exceeds in terms of accuracy between all four algorithms. This time, KNN achieves the lowest amount of accuracy. For a prediction task of classifying between only three labels, the accuracy from the algorithms are rather low in this case. Finally, we also attempted to classify between sex and age range.

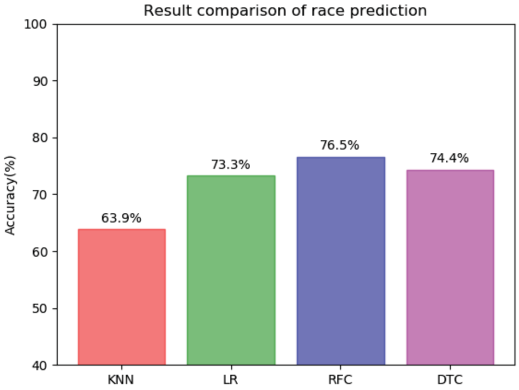

Fig. 3. Comparison between accuracy of four models (2)

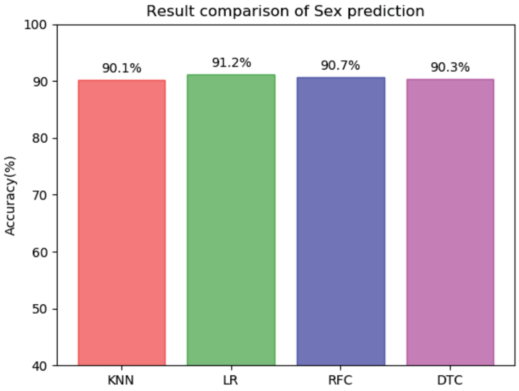

Fig. 4. Comparison between accuracy of four models (3)

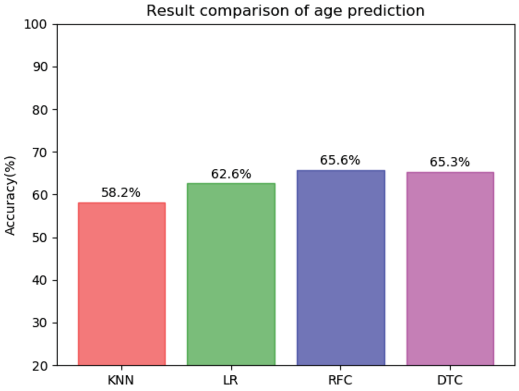

Fig. 5. Comparison between accuracy of four models (4)

TABLE II
MULTIPROGRAM SETS

| | Best and worst accuracy | | | |
|---|---|---|---|---|
| | Best | | Worst | |
| Label | Model | Accuracy | Model | Accuracy |
| Crime Method | RFC | 83.3% | DTC | 80.6% |
| Criminal Race | RFC | 76.5% | KNN | 63.9% |
| Criminal Sex | LR | 91.2% | KNN | 90.1% |
| Criminal Age Range | RFC | 65.6% | KNN | 58.2% |

The results for sex classification were quite good. However, there were only 2 different types of labels for sex so the results were quite understandable. Meanwhile, all the algorithms achieved rather poor results during age range prediction with 65.6% being the highest and 58.2% being the lowest.

In table II, a quite obvious pattern of result is present. In the task of classifying between four different labels, RFC provides the most accurate result in three out of four cases. On the other hand, in three out of four cases, KNN provides the least accurate result. This result gives us an interesting perspective that ensemble classifiers like RFC may provide the most accurate outcome.

## VI. CONCLUSION AND FUTURE WORKS

Our goal of the research was to establish an expandable knowledge that can be used for building machine learning based applications that can reliably output criminal data after giving some victim data and crime information as input. While our current version of research does well for classifying some of the labels such as gender or crime method, there are still lots of improvement needed to be done as there are some obvious weakness of the model. First of all, a criminal cannot be completely identified by just one single attribute. Therefore, multiple attributes are needed to be stacked to create an overall criminal profile. However, when multiple attributes with little errors are stacked, the amount of total error increases by probabilistic theory. Therefore, each of the labels has to be classified very accurately in order to build a successful model. Secondly, when a victim's body is unrecognizable because of burn or some other cause, then data for the proposed system cannot be collected in proper way. Unfortunately, there is no viable solution to this issue for a model like this.

As for future work, the first thing we need to do is to collect more data in order to see if the performance of the classifiers can be improved. In addition to that, the types of features can be experimented as it may well be the case that the current set of features do not fit the labels well enough. Perhaps there are some other important attribute that can provide more information regarding the criminal. Finally, the whole system can be integrated into a database for ease of access and modularity.

## REFERENCES

[1] Rod Gehl and Darryl Plecas. *Introduction to Criminal Investigation: Processes, Practices and Thinking*. Justice Institute of British Columbia, 2017.

[2] Md Abdul Awal, Jakaria Rabbi, Sk Imran Hossain, and MMA Hashem. Using linear regression to forecast future trends in crime of bangladesh. In *2016 5th International Conference on Informatics, Electronics and Vision (ICIEV)*, pages 333–338. IEEE, 2016.

[3] Shyam Varan Nath. Crime pattern detection using data mining. In *2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops*, pages 41–44. IEEE, 2006.

[4] Lawrence McClendon and Natarajan Meghanathan. Using machine learning algorithms to analyze crime data. *Machine Learning and Applications: An International Journal (MLAIJ)*, 2(1):1–12, 2015.

[5] Clifton Phua, Vincent Lee, Kate Smith, and Ross Gayler. A comprehensive survey of data mining-based fraud detection research. *arXiv preprint arXiv:1009.6119*, 2010.

[6] Umair Saeed, Muhammad Sarim, Amna Usmani, Aniqa Mukhtar, Abdul Basit Shaikh, and Sheikh Kashif Raffat. Application of machine learning algorithms in crime classification and classification rule mining. *Research Journal of Recent Sciences ISSN*, 2277:2502, 2015.

[7] Ying-Lung Lin, Meng-Feng Yen, and Liang-Chih Yu. Grid-based crime prediction using geographical features. *ISPRS International Journal of Geo-Information*, 7(8):298, 2018.

[8] Luiz GA Alves, Haroldo V Ribeiro, and Francisco A Rodrigues. Crime prediction through urban metrics and statistical learning. *Physica A: Statistical Mechanics and its Applications*, 505:435–443, 2018.

[9] Raymond E Wright. Logistic regression. 1995.

[10] Leif E Peterson. K-nearest neighbor. *Scholarpedia*, 4(2):1883, 2009.

[11] Xie Niuniu and Liu Yuxun. Review of decision trees. In *2010 3rd International Conference on Computer Science and Information Technology*, 2010.

[12] Roger J Lewis. An introduction to classification and regression tree (cart) analysis. In *Annual meeting of the society for academic emergency medicine in San Francisco, California*, volume 14, 2000.

[13] Carla CM Chen, Holger Schwender, Jonthan Keith, Robin Nunkesser, Kerrie Mengersen, and Paula Macrossan. Methods for identifying snp interactions: a review on variations of logic regression, random forest and bayesian logistic regression. *IEEE/ACM transactions on computational biology and bioinformatics*, 8(6):1580–1591, 2011.