

Machine Learning Approach to Identity Resolution for Criminal Profiling

Hassan Kazemian and Subeksha Shrestha

AI and Data Science Research Group, School of Computing and Digital Media, London Metropolitan
University, London, UK

Abstract - A common dilemma when working on criminal data is that often people manipulate their details to disguise themselves and hide their identities which leads to creating ambiguous and false identities. Deep Neural Network (DNN) is applied to work well on fraudulent and imbalanced data. Two subcategories of DNN, the Multilayer Perceptron (MLP) and Long Short-Term Memory (LSTM) with python libraries such as TensorFlow and Keras are applied to the dataset for the detection of suspects with false identities to assist the process of analytical investigation by law enforcement agencies. Upon application of this approach to the anonymized policing dataset from SPIRIT Project funded by European Union's Horizon, 5 main suspects with false identities were identified out of 23 targets on 39 million records. As working on criminal data is quite sensitive so to avoid data leakage while training and testing the data, K-fold validation has been applied. Furthermore, cascading both MLP and LSTM models on the policing dataset resulted in improved model prediction and accuracy compared to using each model individually. The cascaded model notably reduced false predictions for the recurring criminal patterns such as age-specific trends and most prevalent crime activities.

Keywords: Identity Resolution for Criminal Profiling, MLP, LSTM, TensorFlow, Keras,

1. Introduction

Identity Resolution is an imperative topic that has been researched with formulation of many frameworks and machine learning techniques that have been designed and implemented to achieve most accurate results. Although significant challenges still prevail on the pathway to achieve an optimal result some of which includes difficulty in accessing genuine and high-volumed data, selection of appropriate parameters to build a model and dealing with imbalanced dataset. This paper includes research on resolving those challenges, finding various approach to deal the issues and production of optimal result from the model built. Working with crime detection and prevention requires higher degree of precision particularly when people engage in identity deception to create false identities to mislead the investigation process for various reasons (Fu, et al., 2010).

Deep Neural Network can be efficiently implemented in tackling critical and sophisticated problems even in complex scenarios where higher accuracy is essential so TensorFlow, a machine learning platform that performs with excellence in carrying tensors which are multi-dimensional arrays between the multiple nodes is applied in this project (Barham, et al., 2016). MLP with TensorFlow has been popularly applied to analyse customer behaviour, traveller's history, image recognition, text classification, fraudulent detection, self-driving cars, sentiment analysis and many more (Gan, et al., 2023). Although TensorFlow is popularly applied in various sectors for fraudulent detection but has not been extensively used with crime detection which is the core reason for this research. One of the eminent reasons for fewer research on crime data itself is due to the fact that there is limited access to quality and volumed data, risk of inaccurate outcome resulting in serious penalty for innocent person being prosecuted and setting off the guilty free (Phillips, et al., 2020). Law enforcement agencies constantly seek to use data mining and machine learning approaches to dig deep to extract more information from crime data so patterns can be learnt from past crimes to understand and investigate crime with an aspiration to obtain better results (Kazemian, et al., 2022).

2. Related work

Fraud detection has quite an increasingly critical challenge associated with people sharing more personal information online and conducting transactions through computers, resulting in identity theft, damage and huge financial loss too (Becker, et al., 2010). Working on resolving traditional limitations of inadequate data, resolving inefficient analysis process and possibility of integrating data from different sources such as crime information from surveillance cameras, social attributes to associated from social media, geographical information (Saravanan, et al., 2020). In a study by Dakin, et al., images from Google Street View were processed to identify and extra information on crime, generate location and understand the correlation between features. The paper has some limitation on quality data on detailed

location of the occurrence of crime, size of the data worked and limited features explored (Dakin, et al., 2020). To retrieve underlying information from crime data and also has been beneficial for digital forensics investigators (Jesus, 2011). In a study by Metsker, et al., applied Apache Spark to process the data and more detailed investigation on the crime data with three main categories precisely focused which were civil cases, cases of administrative offences and criminal cases (Metsker, et al., 2019).

In machine learning approaches producing a fair result is also equally important so on research by Pastaltzidis, et al., was carried out on an experimental data, RWF-2000, a large video surveillance data used for detection of violence (Pastaltzidis, et al., 2022). Mask-RCNN algorithm is applied in the study to understand some key points in a video and study the eye, neck, wrists, etc. There are some limitations to this research due to low quality video footage and even limited or absence of videos in some cases. Hybrid modelling has recently been one of acclaimed approaches to improve results while building a ML model and applying it to resolve identity issues. In a study by Al-Sarem, et al., two models namely LSTM (Long Short-Term Memory) and PCNN (Parallel Convolutional Neural Networks) has been cascaded (Al-Sarem, et al., 2021). In this approach the cascaded model was better in detection of rumours related to the COVID-19 pandemic in social media. Particularly for this research, Twitter dataset was processed to detect the fake tweets. In another research by Krishnan & Magalingam, a cascaded model prepared with Recurrent Neural Network (RNN) and Long Short Term Memory (LSTM) which proved to be better than the models were applied individually and had accuracy of over 98% which has improved the model prediction and in-depth analysis in making decision to predict heart related diseases early (Krishnan & Magalingam, 2021).

The model is initially built using TensorFlow a robust platform for machine learning with flexible architecture and provides an opportunity in curating the suitable parameters for activation and optimization functions to obtain the most optimal outcome (Snachez, et al., 2019). Alongside with TensorFlow, Keras which is a high-level API is also implemented to build the model and process the data. Identity resolution is an integral part in crime investigation for the law enforcement agencies and this research demonstrates application of powerful and efficient machine learning approaches with additional features of drop out and L2 regularization to deal with overfitting of the model. So, primarily with TensorFlow, the Multilayer Perception is applied which is a sub-category of DL (Deep Learning) which is a supplementary approach of feed forward artificial neural network that consists of three main layers: input, hidden and output layers. This approach has multiple hidden layers justifying its name being multilayer perceptron where from the input layers data is feed in forward direction to numerous hidden layers finally leading to a predictive output layer (Abirami & P., 2019).

Another sub-category of the Deep Neural Network is LSTM which was also applied on the data using TensorFlow, an approach which consists of recurrently associated blocks each consisting of one or more recurrently connected memory cells having three essential multiplicative units: input, output and forget gates (Deepak, et al., 2021). The presence of forget gate makes this architecture capable of working on handling problems associated with long-term dependencies and making it distinctive from the MLP approach that instead has dense hidden layers (Graves & Schmidhuber, 2005). As both of these approaches produced good results a new model was prepared by cascading both MLP and LSTM architecture to produce even better results. So, this research includes study on application of novel approach of cascading two Deep Neural Network models to improve model accuracy ultimately leading to accurate prediction of false identities and assisting in scaling up prevailing identity resolution approaches.

3. Insights on the Policing Dataset

About the Data

The dataset we are working on is a policing data which is primarily set into two separate datasets one of which has only the details on various crimes and the other dataset has only details of people. The dataset with crime details has all the crucial details such as the first date and time the crime was committed, last date and crime committed, all essential locations details where the crimes were committed which includes district, town, address with postcode, also the nearby police beat number to the crime scene. Additionally, the northing and easting coordinates are also provided to precisely locate the addresses on map. The other dataset consists of personal information of people hence before proceeding the

data, these details are anonymized following the GDPR regulations to protect privacy of personal details. It includes personal details such as forename, surname, date of birth, address and their role in the crime. There is one column which binds both these datasets together which is the “crime reference number”. Each records contains details of every two individuals, where variables with extension “1” and “2” represent the first and second individual respectively. Further analyse is conducted to isolate identical records and discover matching datapoints. Figure 1 shows the combination of the raw crime and personal details of the offenders.

id1	id2	nominal_ref1	nominal_ref2	name1	name2	dob1	dob2	crime_beat_1	crime_beat_2	gender_1	gender_2	role_type_1	role_type_2	ethnicity_1	ethnicity_2	offence_cat_1	offence_cat_2	...	match
0	21738	345	4510110630E	191279580Z		26-11-94	13-08-84			M	M	DEFE	DEFE	white-skinned eu	white-skinned eu	damage/arson	abuse/harassment	...	0
1	350	1433	89519236T	89431558Z		23-03-73	19-02-73			M	M	DEFE	VICT	asian	asian	sexual offence	theft	...	0
2	3278	300	59125492B	89267154Q		14-05-82	01-05-82			M	M	DEFE	DEFE	white-skinned eu	white-skinned eu	theft	theft	...	0

Figure 1: Combining the raw crime and personal details dataset

Two quite crucial features in the policing dataset are the role of people in a crime i.e., whether they are victim or the defendant, and the other one is the various offences committed. The top three crimes committed were theft, murder and harassment. For some crimes such as theft, use of arson, sexual and driving offence the ratio of victims is higher than that of the defendants. In contrary to some crimes such as drug offence, providing false offence, attempted escape and some minor offences have relatively lesser victims recorded. These analysis assists in selection of attributes while building the machine learning model later, and provides a glimpse of which offence is common and what is the difference in terms of percentage for role of crimes committed by the people across various offences. Figure 2 shows in a bar chart the role of people in crime, based on various categories of offences they committed. Figure 2 shows in a bar chart the role of people in crime, based on various categories of offences they committed. To enhance the robustness of the findings and demonstrate the model’s applicability to diverse real-world scenarios, various crime offences are incorporated in the model to simulate different crime scenarios and test its generalizability. By analysing offences ranging from theft and assault to fraud and drug-related crimes, we ensured that the model captures a wide spectrum of criminal behaviours. This approach allows to assess how well the model adapts to different types of offence, improving its predictive accuracy and reliability in practical applications.

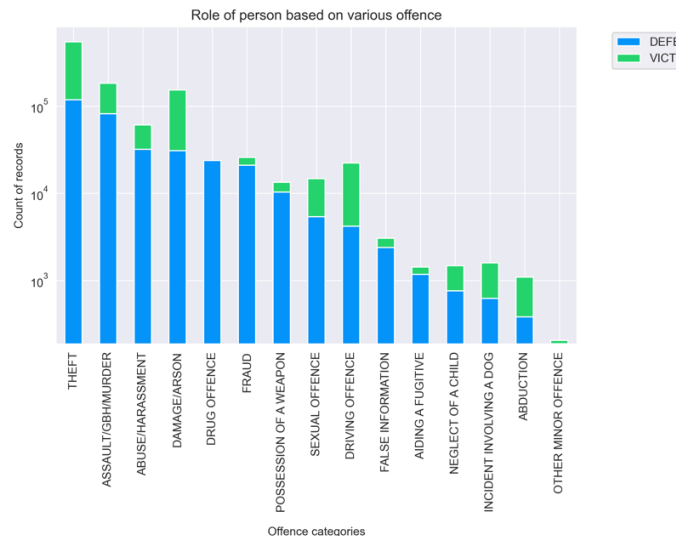


Figure 2: The role of people in crime based on various categories of offences committed

Similarly, the gender of the people in the dataset is also an importance aspect to be considered alongside with the role of a person in crime to understand the spread of these factor across the data. It’s quite evident from the bar graph below that the majority of defendants and victims for both males are quite higher than that of female. Figure 3 depicts the offenders’ gender and their role in crime using bar chart.

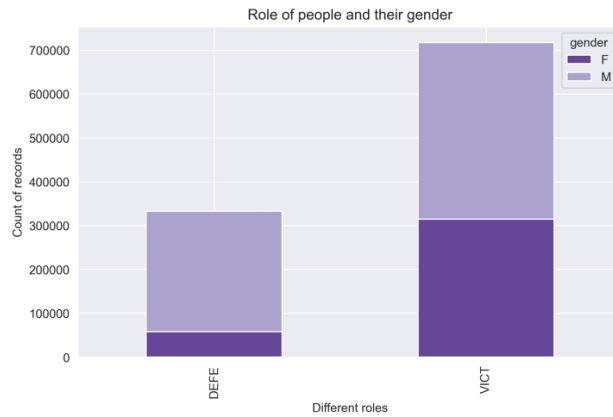


Figure 3: Gender of people and their role in crime

Another attribute is age, which provides year of birth in the policing dataset. This information presents more insights to the data analysis on the average age of the people who committed certain crimes. For instance, the average age of defendants for the theft offence is around 27 and the average age of victims is 43. Figure 4 presents average age of offenders in crime and the registered offences in bar chart.

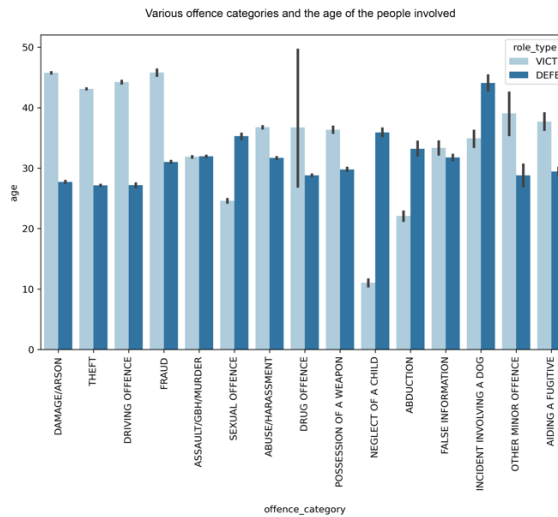


Figure 4: Average age of people in crime and the offences registered

Data Cleaning and preparation

To prepare the data efficiently, over a hundred different categories are sub-categories into just 16 main categories of crime to ease the latter process of analysis. So similar crimes such as rob, burglary, stole, thief, steal, take, make off, etc was clustered into a single group as ‘theft’. Another clustering of group includes classifying the ethnicity based on gender so tracking the criminal profiling becomes more competent for identification of a person. Also, the role of a person in the crime is a crucial attribute to be considered so it's converted into a binary data type which represents victims as 1 and defendant as 0. One of the most crucial implementations on the dataset is consideration of various string-matching algorithms such as Jaro Winkler, Soundex and Levenstein edit distance to compare the first and last name of two strings and compute the similarity score for them (Friendly, 2019). Amongst all the applied string-matching techniques Jaro Winkler has the highest score hence it was applied onto the dataset to match two strings specifically the forename and surname and obtain the similarity score between them. This helps in comprehending the resemblance between each pair

of individuals being compared. Figure 5 demonstrates implementation of Jaro Winkler using Python programming language for the offenders' names.

```
def jaro_winkler(string1, string2, p=0.1, max_l=4):
    # Compute Length of Common Prefix at the start of String
    max_l = max_l if max_l else min(len(string1), len(string2))
    prefix_l = 0
    for char in range(min(len(string1), len(string2))):
        if prefix_l <= max_l and string1[char] == string2[char]:
            prefix_l = prefix_l + 1
        else:
            break

    # Finish Calculations and Return Result
    jaro_sim = jaro(string1, string2)
    winkler_formula = prefix_l * p * (1 - jaro_sim)
    jaro_winkler_sim = jaro_sim + winkler_formula
    return jaro_winkler_sim
```

Figure 5: Implementing Jaro Winkler for the names

Creation of the target dataset and target column

Since the dataset is separately provided, analysis on it becomes difficult so they are merged based on the common factor which is the crime reference where the similarity score of the names, role in crime, date of birth, location and all other factors are considered. Based on the 23 target names provided an individual dataset for all 23 target names are created. Additionally, a target column which is a binary column is created where 1 representing the false IDs manipulated by criminals to hide their identities and the 0 represents the genuine identity and hence 39 million records were formulated. Figure 6 shows a cleaned form of the policing dataset for computer programming.

id1	id2	nominal_ref1	nominal_ref2	name1	name2	dob1	dob2	crime_ref	forename	surname	home_dist	crime_dist	offence	role	year	birthday	...	match
0	21738	345	4510110630E	191279580Z		26-11-94	13-08-94	0	0.000000	0.428571	0.607	0.607	0	1	1.0	0.0	...	0
1	350	1433	89519238T	89431558Z		23-03-73	19-02-73	0	0.430556	0.483333	0.493	0.493	0	0	1.0	0.0	...	0
2	3278	300	59125492B	89267154Q		14-05-82	01-05-82	0	0.455556	0.000000	0.456	0.456	1	0	1.0	0.0	...	0

Figure 6: Cleaned form of the policing dataset

4. Application of MLP, LSTM, and Cascading MLP-LSTM Techniques

Implementing MLP model on the dataset

Intentional manipulation of identities by criminals escalates the risk of misleading investigation and leads to interrogating genuine people and setting criminals free as often criminals tend to hide their personal details to impersonate being an innocent person. The complexity of working on such dataset becomes extremely high as greater amount of accuracy in prediction is the foremost priority while detecting every single fraudulent identity. One of the approaches for better prediction of false identities is figuring out any possible pattern in the way the data have been manipulated (Gordan, et al., 2007). The increase in fraudulent cases related to crime in the recent decades is due to computerization of data as data are vulnerable and process of manipulating of personal details has become much easier.

While implementing machine learning algorithms to the policing dataset, the personal details and attributes related to any crime are studied in detail. To build a model a limited amount of known fraudulent cases are adequate enough to train the dataset and apply that knowledge to learn patterns and trends in the fraudulent cases (United Nations, 2011). With a total of 39 million records in the policing dataset, only about 0.75% of the data are the potential false identities which implies that the training model has very less data to learn the pattern of manipulation of data to distinguish it from genuine records making the policing dataset quite imbalanced. To deal with these imbalances in the data, Synthetic

Minority Oversampling Technique (SMOTE) using additional bias and weight with different parameters are applied to create a more balanced data (Kazemian & Shrestha, 2023). The dataset initially had a significant disparity between the majority (0) and minority (1) classes, with the minority class having very few samples of fraudulent criminal records that were detected. After applying oversampling, the minority class samples are duplicated, which leads to a more balanced dataset and ensured that the model receives sufficient training examples from both classes, reducing bias towards the majority class and improving the overall prediction performance. TensorFlow can successfully be applied to a variety of model to make predictions such as, face or image recognition, text classification, audio recognition and more. In this research, TensorFlow consequently is applied to the policing dataset using a Deep Neural Network based on MLP (Abadi, et al., 2016). Multilayer Perceptron model is applied with TensorFlow and Keras as it has numerous dense multi-layers that are implemented to build the model with different activation and optimization functions that are assigned to improve the performance of the model and equally produce excellent predictions (Rampasek & Goldenberg, 2016).

Amongst many applications of MLP with TensorFlow some of the most famous ones are building financial fraud detection, recommendation for web searches or e-commerce sites and social network feeds (Goldsborough, 2016). The hidden layers are the most crucial aspect of the algorithm to build the model by assigning the appropriate number of hidden layers depending on the type and complexity of dataset (Kazemian & Shrestha, 2023). To increase accuracy of the model different activation and optimization functions are applied along application of various weights and bias. To avoid overfitting of the training dataset which results in a very fixed and biased result dropout function is applied to avoid biased results that is too close to the results from training dataset (Sergeec & Balso, 2018). The dropout values are assigned to the model so that some of the nodes are deactivated for some epoch and they are reactivated in the next step and similarly other nodes are randomly deactivated. This phase of regularization of deep learning by deactivating and reactivating nodes take place to reduce overfitting of the model (Bisong, 2019). Figure 7 outlines the hidden dense layers for the Multilayer Perceptron nodes.

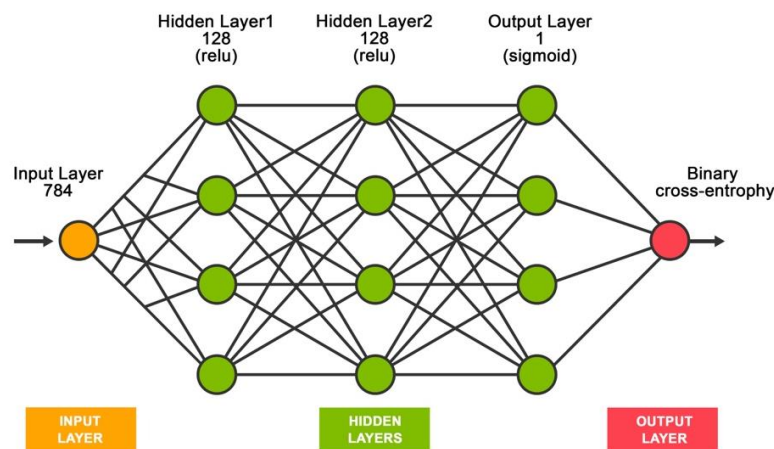


Figure 7: Hidden dense layers for MLP model

The policing dataset contained ethnicity-related data which led to possibility of bias in the data and overrepresentation of certain ethnical groups such as white-skinned European, African-Caribbean and Asian. Therefore, resampling techniques were applied to reduce bias and allow fairness across different demographics by adversarial debiasing and re-weighting. Two other noticeable bias noticed in the dataset were theft and murder, they were amongst the frequently reported crimes. SMOTE was subsequently implemented to maintain a balanced data and to prevent failing to detect other prevailing crimes.

Additionally, adding bias also controls the value activation function by triggering and adapting to the best values. Also, weights are added to improve the model and reduce the amount of loss otherwise. Weights also determine how

much influence the input has on the output node (Pundhir, et al., 2022). Figure 8 demonstrates the sequence of the weights and biases are incorporated to the model.

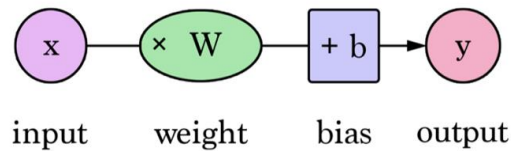


Figure 8: Adding weight and bias to the model (Li, et al., 2020)

Pattern recognition is a common approach when dealing with dataset using MLP as this creates a path to learn from the past prediction made and improve future prediction with each increment in the epoch (Zaccone, 2016). The basic framework of the algorithm includes passing in raw data as input which is then scaled and split into 3 groups to train, test and validate. The model is then built using the training dataset applying numerous functions and creating hidden dense layers that is authenticated with the validation dataset to tune it which was initially separated out. Finally, the model is then tested on the unseen dataset initially assigned as test dataset to check whether or not the predictions made on this dataset is accurate. Figure 9 presents the block diagram of the various computational phases from input data to obtaining the results in the construction of the Multilayer Perceptron model.

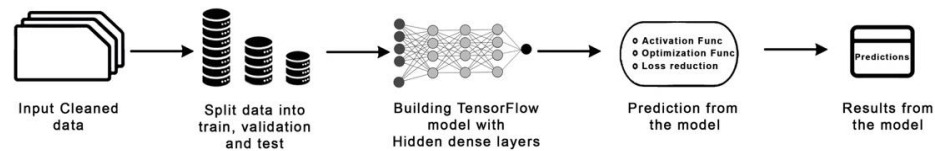


Figure 9: Various phases while building the MLP model

One of the vital aspects of Deep Neural Network model with MLP using TensorFlow that makes it stand out amongst many all-other machine learning approach is its ability to process the data, build the model and obtain the predict at a very high-speed using a powerful API, Keras that helps in building the layers and assigning different functions (Parisi, et al., 2021). As a key part of neural network design, the selection of appropriate activation function from various available choices is essential to make the computation easier and accelerate the process of building the model, ReLU a type of activation function is selected in this case. Additionally, to improve the speed and performance of the training model optimizer functions are also assigned which for this model is Adam optimization function to decay the average gradient value and obtain optimal result swiftly with less parameters to tune.

Implementing LSTM into the model

Another Deep Neural Network architecture Long-short Term Memory (LSTM) based on Recurrent Neural Network has also been implemented. LSTM is applied because of its unique feature to remember long term dependencies and has three main gates and a memory cell state (Kumar, et al., 2019). The three main gates embedded in LSTM is the input, output and forget gate. A fraud detection model is prone to overfitting so various bias and activity regularizers were implemented (Ullah & Mahmoud, 2022). In LSTM, the main and primary stage is to discover the information cell state that will be discarded which is through the medium of the forget gate. The forget gate layer is a sigmoid layer where the output from the previous LSTM block and input from the current timestamp is examined which returns values between 0 and 1. From the outputs, 1 is the values that represents the cell state to retain and 0 represents those cell states that can be discarded.

Then the next sigmoid layer which is the input gate layer is initiated which provides additional information that aids in updating values and the tanh layer creates a potential nominee value for the current state of LSTM. The final step is working on the current cell state to determine the output which begins with a sigmoid layer that is initiated with the values received after which tanh function is applied which is multiplied by the outcome from the sigmoid gate and finally

the values to return are selected (Feng, et al., 2019). So, this process continues which helps in determining which cell state should be retained and which can be eliminated. To get a clear understanding of LSTM, its architecture which begins with forget gate where essential information is retained while unnecessary information is discarded from the previous cell state, then the relevant information is passed to the input layer where selective update is carried out. Finally, the hidden states with appropriate sigmoid and tanh layer are formulated in the embedding layer which then leads to classifying the data into false or genuine identity and producing the results in the output layer. Figure 10 shows the layout of Long-Short Term Memory model outlining Input Layer, Added weights, Hidden Cell State, Additional network features, Embedding layer, and finally Output from LSTM network.

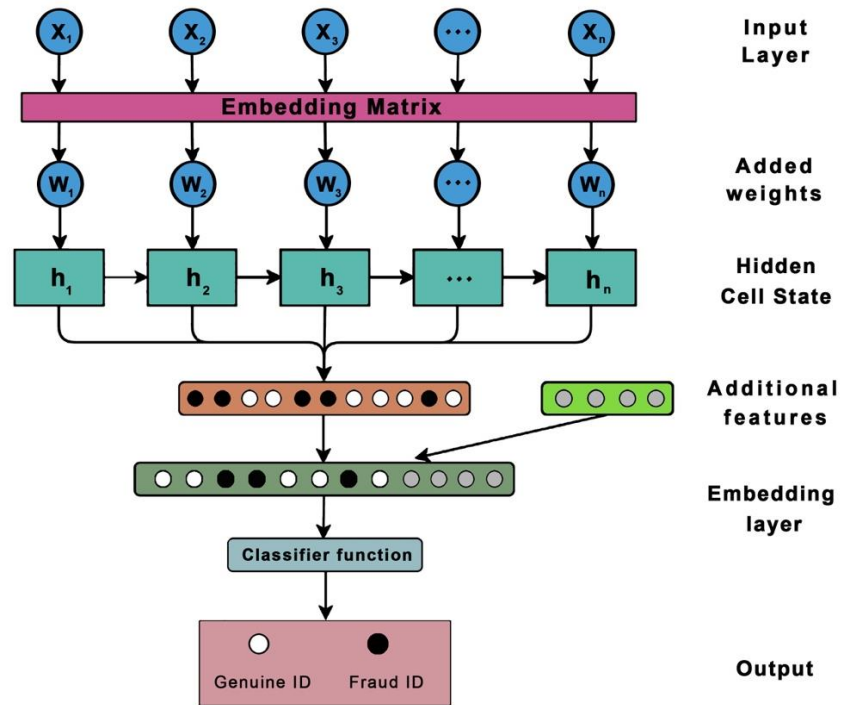


Figure 10: Architecture of LSTM model

Cascading MLP with LSTM model

A hybrid Deep Neural Network model has been designed to combine Multilayer Perceptron and recurrent neural networks onto the policing dataset to improve the rate of prediction and conduct novel research on crime datasets. Multilayer Perceptron on one hand has the competency to work on complex and sensitive dataset with multiple hidden layers and finest activation and optimization function whereas Long-Short Term Memory on the other hand can perform efficiently by resolving long-term dependency issues. Both of these Deep Neural Network models upon application on fraud detection and identity resolution individually has improved the prediction so cascading both the model was initiated to obtain a better predictive model. Another distinct reason to cascade both the models was because MLP produced less inaccuracy for false negative values whereas LSTM produced less false positive values so when both models are combined that could produce minimal false positive and false positive results which ultimately leads to producing high accuracy in the predictions made. Figure 11 depicts the set of parameters for building the cascaded Multilayer Perceptron and Long-Short Term Memory model.

```

print('model started')
model = tf.keras.Sequential([
    tf.keras.layers.Dense(512, activation='relu', input_shape=(X_train.shape[1],)),

    keras.layers.Dense(128, activation='relu'),
    keras.layers.Dense(128, activation='relu'),
    keras.layers.Dropout(0.5),
    tf.keras.layers.Reshape((-1, 128)),
    tf.keras.layers.LSTM(32, activation='tanh', return_sequences=True),
    keras.layers.Dropout(0.5),

    tf.keras.layers.Dense(1, activation='sigmoid')
])

print('model stopped')

print('model compiling')

# Compile the model
model.compile(optimizer='adam',
              loss='binary_crossentropy',
              metrics=['accuracy'])

```

Figure 11: Parameters set for building the cascaded model

Fraudsters try to conceal and blend themselves into the environment to remain unnoticed and try intentional measures of manipulating records. Hence, a hybrid model is built, as the fraud detection system needs improvements (Baesens, et al., 2021). The important part of this model is that both LSTM and MLP are built implementing Keras, but both models have separate functions and parameters. Essentially the model is built initiating TensorFlow and Keras following the architecture of MLP with numerous hidden dense layers. In addition to this rectified linear unit (ReLU) activation function is applied to reduce gradient problems such as weight and bias produced in the initial input layers. Since the preliminary initial layers are crucial while building the model and plays a key role in the overall model accuracy and the phase of building the entire artificial neural network, ReLU an activation function is implemented to minimize such issues with every increment in the epoch in the data training session. Then the LSTM architecture is introduced to the model with embedding layers, forget, input and output gates. Besides MLP with TensorFlow and LSTM model, other classification models such as SVM (Support Vector Machine), KNN (K-Nearest Neighbour) and Naïve Bayes were also applied to determine predictive accuracy and effectiveness (Kazemian & Shrestha, 2023). The objective was to identify the most suitable models for cascading in the subsequent stages of the analysis. Additionally, efficient neural architecture techniques such as early stopping, shared weight and reduction of redundant computations were implemented to streamline the network structure, ensuring that both MLP and LSTM components were optimized for minimal computational overhead. Table 1 shows the accuracy of MLP with TensorFlow, LSTM, KNN, Naïve Bayes, and SVM models based on ROC curve results.

Table 1: Different models results based on ROC curve (Kazemian & Shrestha, 2023)

Model	ROC score	Execution Time
MLP with TensorFlow	99.99 %	1 hr 40 mins
LSTM	99.99 %	8hrs 46 mins
KNN	99.94 %	6 hr 48 mins
Naïve Bayes	90.61 %	5 mins
SVM	85.69 %	6 hrs 32 mins

Hyperparameter tuning and optimization is also initiated into the model to improve the accuracy of the predictive model, reduce the overall loss when fitting the data into the machine learning model. These functions ensure less memory is consumed and higher accuracy of the model is attained so the phase to build the model can be executed in reduced time. A lot of other functions are also implemented such as the drop out where arbitrary neurons are ignored while training the data to avoid overfitting of the model. K-fold cross validation is also considered to prevent leakage between training and

test data. After all these functions and parameters are adjusted on the model, the hybrid model is built to predict the data and test the accuracy of the new model built. Since this model implements both MLP and LSTM it comparatively takes longer execution period than when the models were built individually. Model pruning was used to remove redundant weights and neurons, reducing the model size without significantly affecting performance. Quantization was implemented to lower precision computations, enabling faster execution while maintaining acceptable accuracy. To specifically address the computational overhead, techniques such as batch size optimization and learning rate scheduling were utilized to improve efficiency. Additionally, parallel processing has been employed to speed up the execution of the models, and to ensure the hybrid model remains feasible for practical use despite of its complexity. To measure the accuracy of the model and evaluate outcomes produced, many evaluation metrics were applied. While cascading both of the models, numerous attributes were reconsidered and tuned based on the outputs received, each alteration performed were carefully observed and were then finally implemented to the cascaded model.

5. Simulation Results for Identity Resolution

In the process of predicting fraudulent data, sophisticated string-matching techniques are applied to prevent, detect and predict fraudulent records in the dataset (Netowl, 2025). Different string-matching techniques were considered and applied to the dataset, including three edit-based methods such as Levenshtein edit distance, Hamming distance, and Jaro-Winkler as well as token and phonetic-based techniques. Among these, Jaro-Winkler is selected as the optimal technique due to its effectiveness in handling minor spelling variations and transpositions, making it particularly suitable for identity resolution in criminal fraud detection. The target variables are compared independently with all the other records and Jaro Winkler is applied to calculate the similarity score between every two strings analysed and compared. The home and crime distance of every two records compared were evaluated to observe whether it is greater than the range of a certain kilometres where different values such as 25, 50, 75 and 100km were applied. However, the best result was obtained when the radius was set to 50 km. During the calculation those records having distance greater than 50 km is represented as 1, and for results less than 50 km, it is not set to exact zero but is assigned in decimal value to acknowledge its closeness to 1.

Depending on the role of a person in the crime, if they are the victim in a particular crime then they are assigned as 1 and if they are the defendant then 0. Before the final comparison, another important feature, the types of offence in a crime are also evaluated, the various categories of offence are clustered together as there are over 100 different offences but there are some offences that are similar in some context, henceforth the offence categories were reduced to just 16 main categories to make the analytical process easier and efficient. This categorization of offence is to avoid similar offences being identified as dissimilar ones with the presence of a minor variation in the category which ultimately eases tedious execution phase. Finally, the target column is assigned as a binary value where 1 represents the two strings for comparison are false identities created by the criminals and 0 indicates that there is no manipulation done on the data and these identities are genuine. Figure 12 compares string-matching technique and calculated distance for every two strings.

	crime_ref	forename	surname	home_dist	crime_dist	offence	role	year	birthday	match
New_ID										
0	0	0.000000	0.428571	0.607	0.607	0	1	1.0	0.0	0
1	0	0.000000	0.000000	0.559	0.559	0	0	1.0	0.0	0

Figure 12: Comparing string-matching technique and calculated distance for every two strings

The architecture of Artificial Neural Network (ANN) is extracted from the structure of biological neuron similar to human brain for stimulation of neurons to form a network of dense hidden layers used to build complex predictive models (Alaloul & Qureshi, 2020). In ANN, the neurons are interlinked in such a way to make the model capable of handling large volume of data using the input layer, processing further with the dense hidden layers and finally producing an output as prediction. Assigning a specific number of neurons as the hidden dense layer is a very crucial aspect while

building the MLP model with TensorFlow as the neurons are set based on the exponential power of 2 beginning from 2 up to 512. To improve the accuracy of prediction not just a single layer but multiple combinations of dense hidden layers are implemented. As there is no explicit rule or strategy to assign the dense layers, different combination of hidden dense layers is applied and tracking of the accuracy of each model built and predictions made are noted and the optimum combination is henceforth selected (Table 1). Table 2 presents the False Negative and False Positive values for each selected dense hidden layer.

Table 2: Demonstration of the False Negative and False Positive values for each selected dense hidden layer

Dense Layer	FP	FN
512-128-128-128-1	218	262
512-128-128-1	238	123
512-128-64-1	593	247
512-512-64-1	695	284
512-64-32-1	631	164
512-1	869	1320
256-1	993	1045
128-1	317	3773
64-1	741	2085
32-1	814	2642
16-1	317	3773
8-1	205	4555
4-1	218	4819
2-1	223	6422

One of the best combinations after multiple trial and error approach was assigning 512-128-128-1 where initially 512 input values are inserted, then a double 128 hidden dense layer is assigned and finally 1 optimized output layer is set. The proximity of obtaining an optimal prediction from the model is by training the data, adding weights to the network and tuning various parameters for loss reduction and addressing overfitting issues. In the process of constructing the model to refine the training data epochs are introduced to learn and finally make the predictions. This process sometimes leads to creating an overfitted model which produces predictions biased towards one value so to address this issue drop out functions and early stop on epochs are assigned. Figure 13 plots the computational result for various accuracy and loss graph with increment in the epoch.

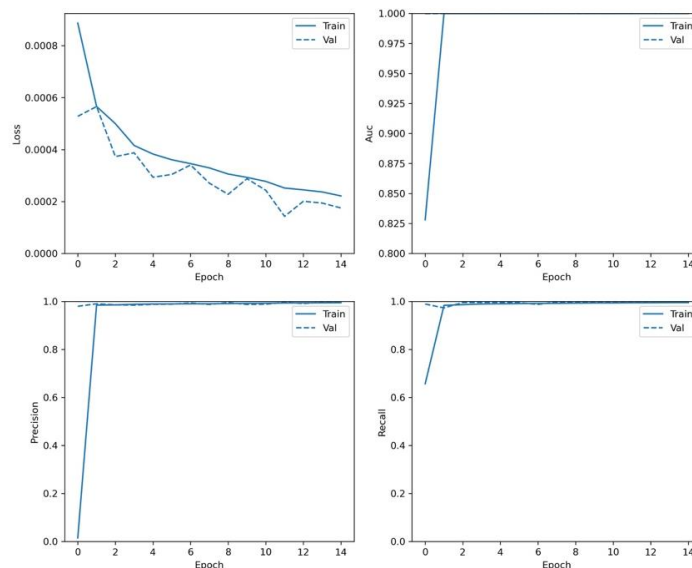


Figure 13: Various accuracy and loss graph with increment in the epoch

Another essential aspect to be considered is the distance to be set for every two compared records where the calculation of the home and crime distance is set to a certain location radar to know what their actual distance range to the potentially manipulated location is. For every two strings compared the distance between the northing and easting point is calculated for both home and crime distance. Euclidean distance formula is applied, and 4 different distance values are considered with the interval of 25km each. One of the optimum results is populated when the distance is set as 50km where any distance greater than 50 km is considered as 1 and less than 50 km is not directly assigned to 0 but with 3 decimal place to observe its closeness to being assigned to 1 (Table 2). Table 3 below demonstrates the results from the built model when different distances were evaluated.

Table 3: Demonstration of the results from the built model when different distances were evaluated

Home & Crime distance compared	FN (False Negative)	FP (False Positive)	Precision (While the ML model is built)
25km	179	505	99.15%
50km	123	238	99.59%
75km	162	704	99.85%
100km	132	328	99.44%

For further approval and selection of the optimum value for crime and home distance, additional measures to compare the correct and incorrect predictions were carried out. Similar to the previous results, the least number of incorrect predictions was found when the distance was set to 50 km, it predicted only 36 incorrect records in approximately 13 million test data from the total of 39 million records which indicates quite a minimal rate of inaccuracy in the model. Table 4 shows the incorrect predictions made by the model based on different set distances.

Table 4: Demonstration of the incorrect predictions made by the model based on different set distances

Distance between compared strings (Home & crime distance)	Incorrect predictions (Outcomes after building the model)
25km	79
50km	36
75km	69
100km	45

Accuracy of a model is a very crucial aspect that determines the performance of a model and particularly when working with sensitive crime dataset it becomes more important to have a higher accuracy rate. Similarly, the precision score is another essential measure that analyses the correctly predicted positive values out of all the predictions whereas the recall score is similar to sensitivity and keeps track of positive instances (Davis & Goadrich, 2006). Another evaluation measure applied is the AUC score which defines that higher the AUC score better the model prediction so when classifying the target variables which in case of the policing dataset is to identify between fake and genuine IDs, the higher AUC score it validates that the model has correctly predicted majority of the labels (Huang & Ling, 2005). Loss in a model is considered as a penalty for an incorrect prediction made so when a model predicts nearly perfect it has relatively less loss compared to a badly predicted model. The rate of loss is an important aspect that should be considered in the phase of building a model and it should be as minimal as possible.

The results presented in Table 5 provide a comprehensive evaluation of the models using various performance metrics, including accuracy, precision, recall, F1-score, and AUC-ROC. Accuracy gives an overall measure of correctness, but it may not fully reflect model performance in cases of class imbalance, so further evaluations are performed (Zhao & Teng, 2025). For crime investigation and identity resolution, precision and recall are critical. The Cascaded model (TensorFlow & LSTM), with high precision (0.9985) and recall (0.9988), demonstrates an effective identification of true positives while minimizing false positives (Evidently AI Team, 2019). The F1-score of 0.9986 further confirms this balance. Additionally, the AUC-ROC score of 0.9999 shows excellent model ability to distinguish relevant data, vital for accurately, linking identities in crime-related tasks (Dash, 2022).

Table 5: Evaluating model's performance

Model	Accuracy	Precision	Recall	F1-score	AUC-ROC score
Cascaded model (TensorFlow & LSTM)	0.9985	0.9985	0.9988	0.9986	0.9999
MLP with TensorFlow	0.9900	0.9956	0.9637	0.9794	0.9999
LSTM	0.9975	0.9807	1.0000	0.9903	0.9999
KNN	0.9930	0.9923	0.9965	0.9944	0.9994
Naïve Bayes	0.8125	0.7746	0.9917	0.8698	0.9061
SVM	0.8450	0.8800	0.9850	0.9290	0.8569

A confusion matrix consists of two-dimension results one of which has the actual values and the other have the predicted values (Haghighi, et al., 2018). The lesser the value of false negative (actual positive values that are predicted as negative) and false positive (actual negative values that are predicted as positive) values predicted by a model, the better and precise the model would be as it could consist of least incorrect predictions. For an in-depth evaluation, confusion matrix which has 4 main values each having an individual importance which is outlined below particularly considering the policing dataset. Out of over 25 models built applying various parameters, functionalities, python libraries and different machine learning models, the top 3 results have been considered to illustrate the concept of confusion matrix. The first confusion matrix (denoted as 14 'a') is the result from the model built implementing only MLP with TensorFlow and Keras under various Deep Neural Network models and application of k-fold validation to avoid data leakage between train and test data. The second confusion matrix (denoted as 14 'b') was obtained by implementing only the LSTM model on the policing dataset. And the final confusion matrix (denoted as 14 'c') was built cascading both the Deep Neural Network models MLP and LSTM.

Figure 14 outlines confusion matrix results for the 3 different ML models. Upon careful analysis on all 3 results on various aspects, the third confusion matrix (denoted as 14 'c') produced the least number of false negative and false positive values amongst all 3 models which is the core aspect of building and predicting the model and conducting this research. In the cascaded hybrid model of MLP and LSTM too the potential manipulated identities are predicted accurately with highest accuracy so a detailed description of all the 4 components of a confusion matrix is explained below (in context to figure 14c).

True Negatives : 13020359
 False Positives : 138
 False Negatives : 0
 True Positives : 102069

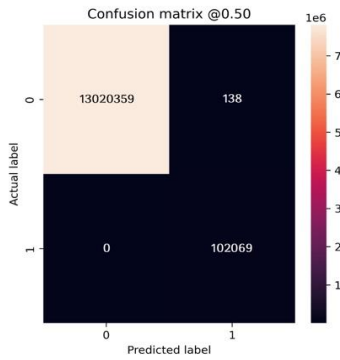


Fig. (a) MLP model

True Negatives : 7814058
 False Positives : 67
 False Negatives : 5018
 True Positives : 54397

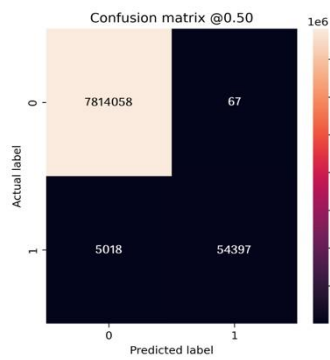


Fig. (b) LSTM model

True Negatives : 13020495
 False Positives : 2
 False Negatives : 0
 True Positives : 102069

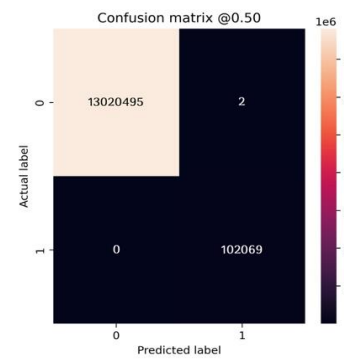


Fig. (c) Cascading model with MLP & LSTM

Figure 14: Confusion Matrix results for the 3 different ML models

Upon further analysis only on the False Negative and False Positive values, in the first confusion matrix result from MLP model the False Positive is 138 but in the LSTM model it is further reduced to 67. So, the cascading of MLP model with the LSTM was initiated which further reduced the prediction of False Positive values to just 2. Concurrently the False Negative values in MLP model was 0 but in LSTM model had about 5018, which was initially alarming but upon cascading the functionality and architecture of both the DNN models the final outcome of False Negative values was again reduced to 0. Although MLP model had False Negative values predicted as 0 but it had a higher False Positive value compared to LSTM so the cascading of these two models were formulated. If the MLP model was deployed alone then there would be many cases where genuine people would be questioned, and investigation process could be time consuming because of the inaccuracy it had. But with the reduction in False Positive cases to just 2 records out of 13 million test records it validates that the predictions made from the model is quite accurate.

To achieve this accuracy in the model several features and functions were applied to aid in tuning of various parameters. One of the approaches adopted to improve the model was to implement drop out value that is used to arbitrarily deactivate neurons in a single training set producing the output as 0 irrespective of the original value to obtain an unbiased prediction in the end. Additionally assigning of this drop out value reduces the chances of overfitting which is a condition when the model has been perfectly fitted to the training model which upon testing gives biased results and can't exhibit the true essence of a predictive ML model. As these are tedious task so to improve the speed and performance of the model optimizer functions are essential which in this model is assigned to Adam (Adaptive Moment Estimator) which focuses on minimizing loss while building a model (Tato & Nkambou, 2018).

The model is not just built by executing the algorithms and functions created once to train the dataset, but this phase is iterated a number of times as the model requires to learn the pattern with every increment in the epoch, so a certain value is set which in case of the policing dataset was set to 100 epochs in total. So, basically the procedure of building, training, and fitting the data to the criteria set while initialising the model would be repeated 100 times. But during this phase having the same result for a continuous period would not help train the data to learn anything and improve but would lead to an over fitted model which is biased towards one of the outcomes. To avoid this scenario a method called early stopping is applied which implies that the epoch would be halted if the same result persists for several iterations and if there is nothing new for the model to be trained on. So, while building the model, although the initial epoch has been set to train the model at 100, it stops for instance at 30 or 31st epoch to produce the optimal result from the model instead of completing the entire 100 epoch.

This approach has another benefit too which is that it would save huge amount of execution time when the model is trained by repeating the process repeatedly for about 100 times. As early stopping is introduced to the model which stops the model when there is no new pattern to learn from for the training dataset this eventually avoids overfitting and

reduces the runtime significantly (Hastomo, et al., 2021). Finally, to reduce loss while building the model, binary cross entropy is selected as we need to get the outcome as a binary value which is 1 for false and manipulated identities whereas 0 represents genuine identities. The cascaded model was successful in predicting out all the 5 main suspects in the policing dataset out of over 39 million records, a closer analysis on the results was carried out to discover pattern in the manipulation of records. Firstly, the age band of the suspects were investigated which was mainly between the age group of 22 to 32 which indicates that some of the crimes were committed at a young age too and some around mid-thirties. Figure 15 depicts in bar chart the details for 5 suspects successfully predicted from the cascaded Multilayer Perceptron and Long-Short Term Memory model.

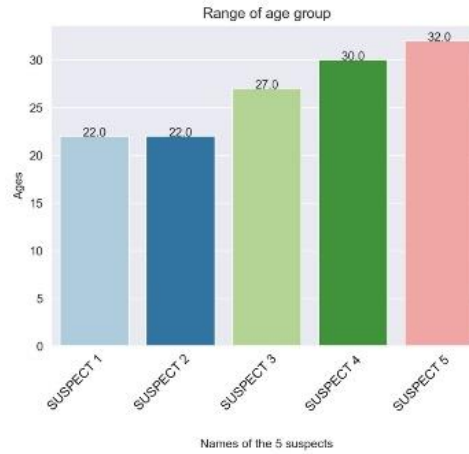


Figure 15: Details for 5 suspects successfully predicted from the cascaded model

After further investigation many crucial insights were obtained, mainly the suspects were male, and their ethnicity was white skinned Europeans. The crimes that were mostly committed were either theft of motor vehicles or burglary dwelling. A notable pattern observed in the manipulation of identities was the recurrence of crimes primarily between 2015 to 2018, suggesting a repeated trend in offences during this period. Figure 16 presents clear predictions from the cascaded Multilayer Perceptron and Long-Short Term Memory model.

	name	gender	age	offence	role_type	ethnic_group	crime_first_committed	crime_last_committed	distance between crime scene and home distance
0		M	30	THEFT FROM MOTOR VEHICLE	DEFE	WHITE SKINNED EUROPEAN	25-Apr-17	-	5mins
1		M	27	THEFT FROM MOTOR VEHICLE	DEFE	WHITE SKINNED EUROPEAN	17-Sep-15	01-May-17	6 mins
2		M	22	THEFT FROM MOTOR VEHICLE	DEFE	WHITE SKINNED EUROPEAN	18-Sep-15	24-Feb-18	8 mins
3		M	22	BURGLARY DWELLING	DEFE	WHITE SKINNED EUROPEAN	18-Sep-15	11-Mar-18	47mins
4		M	32	TAKE MOTOR/VEH W/O OWNER CONSENT	DEFE	WHITE SKINNED EUROPEAN	18-Sep-15	22-Jun-18	47mins

Figure 16: Insights from the cascaded model predictions

Figure 17 demonstrates all suspects have been predicted correctly by the cascaded Multilayer Perceptron and Long-Short Term Memory model. Upon evaluating the cascaded model, it had the highest accuracy and the least rate of inaccurate predictions compared to the MLP and LSTM models built individually so further computations were applied using the cascaded model for prediction. The cascaded model was successful in predicting all the 5 suspects and this has helped to reduce the tedious task of Law Enforcement Agencies (LEAs) to investigate and find the suspects swiftly lot of

the large datasets. The higher accuracy is particularly important for the policing dataset since the offence committed are horrendous and not being able to detect the manipulated records and resolve the identity crisis would be a serious issue.

	New_ID	id1	id2	nom_ref1	nom_ref2	name1	name2	dob1	dob2	name_dob1	...	home_dist	crime_dist	offence	role	year	birthday	match	Prediction
Unnamed: 0	61	8570040	18982	7081	51361681U	51189179681U		26-05-92	26-05-92		...	0.975	0.975	1	1	1.0	1.0	1	1
	235	304	21738	25704	4510110630E	4570903910630E		26-11-94	26-11-94		...	0.942	0.942	0	1	1.0	1.0	1	1
	358724	6960104	454987	240572	16589105T	16915871105T		16-03-90	16-03-90		...	0.978	1.000	1	1	1.0	1.0	1	1
	32400	10033117	330565	331276	279192762	2711007412762		23-11-99	23-11-99		...	0.936	0.936	1	1	1.0	1.0	1	1
	32509	10033394	330565	345821	279192762	2711007412762		23-11-99	23-11-99		...	0.911	0.911	1	1	1.0	1.0	1	1

5 rows x 24 columns

Figure 17: All suspects predicted correctly by the cascaded model

As the cascaded model had a very minimal rate of inaccurate predictions when evaluated implementing confusion matrix and other evaluation measures, so further research on the outcomes for inaccurate predictions was carried out. As there were 0 false negative values in the cascaded model, so the model was efficient in uncovering all the manipulated records but as the false positive values on the test data was about two, so some genuine records have been considered as false or manipulated ids. Upon critical analysis on the results obtained it was due to various similar factors such as their same date of birth, similar category of offence, same role in the crime and a very near crime location which is shown in the figure below where they are '0' or genuine records and they have been predicted as '1' or manipulated records. Figure 18 shows minor inaccurate predictions from the cascaded Multilayer Perceptron and Long-Short Term Memory model.

	New_ID	id1	id2	nom_ref1	nom_ref2	name1	name2	dob1	dob2	name_dob1	...	home_dist	crime_dist	offence	role	year	birthday	match	Prediction
	2	32566965	213039	373241	4936442A	1112288923C		12-01-85	12-01-85		...	0.608	0.608	0	1	1.0	1.0	0	1
	14	2005	21738	179545	4510110630E	109535242E		26-11-94	26-11-94		...	0.669	0.669	0	1	1.0	1.0	0	1
	35	32535693	594473	66561	40864456L	401215198H		29-11-05	29-11-05		...	0.824	1.000	1	1	1.0	1.0	0	1

3 rows x 24 columns

Figure 18: Minor inaccurate predictions from the cascaded model

6. Conclusion

The research introduces a novel approach of combining two Deep Neural Network models, Multilayer Perceptron and Long Short-Term Memory for identity resolution. This methodology includes testing of various parameters beginning with selection of suitable number of dense hidden layers, clustering of groups to avoid ambiguity, setting different location radar for crime and home distance to obtain the most optimum result. The research includes identity resolution on anonymized policing dataset which is a part of the SPIRIT project funded by the European Union's Horizon. The methodology has successfully identified 5 false identities from 23 main targets out of over 39 million records. The aim and objectives set at the beginning of the project had been successfully achieved with this research which had dealt with numerous issues prevalent in the field of identity resolution. This had been obtained with application of quality data, application of ML techniques on large volumed data and cascading ML models to produce better predictions. Deep analysis on various entities of the dataset had been carried out to understand the patterns adopted by criminals to manipulate their details, measures taken to hide their true identity and ways criminals mislead the investigation process.

The research also focused on applying various machine learning algorithms to resolve identity issues and upon further investigation a novel approach of combining two Deep Neural Network models, Multilayer Perceptron with TensorFlow and Keras and Long Short-Term Memory had been applied. This research included testing of various parameters beginning from selection of suitable number of dense hidden layers, clustering of groups to avoid ambiguity and setting different location radars for home and crime scene had been implemented on the policing dataset. Since one of the DNN models Multilayer Perceptron had been efficient in prediction with lower number of False Negative values and the other model Long-Short Term Memory had been proficient with prediction particular with reduced number of False Positive values so cascading of MLP with LSTM had been initiated. Upon creating the cascaded model, the prediction had been greatly improved as both the False Negative and False Positive values had been greatly reduced. This ultimately helped in achieving an optimum result that helped in distinguishing false and manipulated identities accurately and helped in investigation procedure for law enforcement agencies in identification of potential false identities in a swift and accessible manner.

The predictions made from the model helped in making the tedious investigation procedure much easier in terms of computation and accessibility for the Law Enforcement Agencies to have an in-depth analysis on the dataset and the entities associated with the fake IDs produced. The model also helped in investigation on the policing dataset to observe the outcomes from the individual models and analyse on how combining both the models had been beneficial for the prediction of the crime records and assisted in resolving identity issues. To achieve this goal, validation of training dataset had been performed using k-fold validation one of the most efficient measures to ensure there had been no data leakage and to ensure higher accuracy of the model. Then additionally parameters had been adjusted, and application of various optimization, activation and loss functions had been applied on the model to improve the accuracy and prediction. The model had been built with powerful ML library such as TensorFlow and high-level APIs like Keras which extensively improved the model predictions. These 5 suspects had mostly been involved in similar crimes which were either theft of vehicles or burglary dwelling, and their age group had been mainly between 22 to 32. Another interesting insight obtained from the data had been that these manipulation of personal details and identities mostly had been of criminals who had previous committed similar crimes in the past.

The research primarily focuses on a policing dataset as part of the SPIRIT project funded by the European Union's Horizon, and it successfully identifies false identities within law enforcement records. While the methodology demonstrates effectiveness in crime-related identity resolution, its application to other datasets or crime types has not been explored in this study. Future research can explore extending the model's applicability beyond policing, including sectors like finance, healthcare and more. Further studies could assess the model's performance on diverse datasets to expand its applicability and use across different sectors. The findings of this research provide valuable insights into criminal identity manipulation and enhance investigative processes for law enforcement agencies. The model's ability to detect false identities can significantly aid in crime prevention and efficient resource allocation. However, considering the socio-technical aspects, it is essential to evaluate the ethical implications of automated crime analysis, such as potential biases in data collection and model predictions. The real-world impact extends to policymaking, where such models could provide informed decisions on fraud detection strategies and law enforcement protocols. Future work should also assess how law enforcement agencies can effectively integrate this model into their workflows, while ensuring transparency and accountability in its application.

Future research would include study on possible ways to reduce the execution time for building the model alongside with research on improving accuracy of the model with the reduced False Negative and False Positive values. Additionally further study on other possible approaches to cluster ethnicity, gender, location, etc could be attained while getting the target variable and enhance the analysis of the policing dataset. As there is no generic rule in selection of values for hidden dense layers or embedded and LSTM layers so further different approach could be researched in attempts to discover an easier approach on find the most suitable values while building the model. This cascaded model of MLP with LSTM was only tested on policing dataset as a part of the project research but further investigation to broaden the scope of the model could be done by application of the model build across other sectors too such as in finance, healthcare, etc.

Acknowledgements

This work was supported by the European Union Horizon SPIRIT project under Grant Number 786993. We would like to thank all our SPIRIT Project partners who provided us recommendation and feedback that greatly assisted in the improvement of this research.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability

Due to European Union Horizon data confidentiality and GDPR, the names in the dataset are anonymized and is only available under secure organization's fileserver with encryption. But upon request a sample of the dataset could be made available with the names and other personal details would be removed to abide by the confidentiality agreement.

REFERENCES

- Abadi, M. et al., 2016. Savannah, GA, USA, Google Brain, pp. 265-283.
- Abirami, S. & P., C., 2019. Chapter Fourteen - Energy-efficient edge based real-time healthcare support system. *Advances in Computing*, 117(1), pp. 339-368.
- Alaloul, W. S. & Qureshi, A. H., 2020. *Dynamic Data Assimilation*. 2nd Edition ed. London: IntechOpen.
- Al-Sarem, M. et al., 2021. A Novel Hybrid Deep Learning Model for Detecting COVID-19-Related Rumors on Social Media Based on LSTM and Concatenated Parallel CNNs. *MDPI*, Volume 11, pp. 2-170.
- Baesens, B., Hoppner, S. & Verdonck, T., 2021. Data engineering for fraud detection. *Elsevier*, pp. 1-13.
- Barham, P. et al., 2016. *TensorFlow: A System for Large-Scale Machine Learning*. Savannah, GA, USA, Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI '16)..
- Becker, R. A., Volinsky, C. & Wilks, A. R., 2010. *Fraud Detection in Telecommunications: History and Lessons Learned*, s.l.: Technometrics.
- Bisong, E., 2019. Regularization for Deep Learning. In: *Building Machine Learning and Deep Learning Models on Google Cloud Platform, 2019*. s.l.:Apress, pp. 415-421.
- Dakin, K. et al., 2020. Built environment attributes and crime: an automated machine learning approach. *Crime Science*, 9(12), pp. 2-17.
- Dash, S., 2022. *Understanding the ROC and AUC Intuitively*. [Online] Available at: <https://medium.com/@shaileydash/understanding-the-roc-and-auc-intuitively-31ca96445c02> [Accessed 18 February 2025].
- Davis, J. & Goadrich, M., 2006. *The Relationship Between Precision-Recall and ROC Curves*. Pittsburgh, PA, Appearing in Proceedings of the 23rd International Conference on Machine Learning.
- Deepak, G., Rooban, S. & Santhanavijayan, A., 2021. A knowledge centric hybridized approach for crime classification incorporating deep bi-LSTM neural network. *SpringerLink*, Issue 80, p. 28061–28085.
- Evidently AI Team, 2019. *Accuracy vs. precision vs. recall in machine learning: what's the difference?*. [Online] Available at: <https://www.evidentlyai.com/classification-metrics/accuracy-precision-recall> [Accessed 18 February 2025].
- Feng, M. et al., 2019. Big Data Analytics and Mining for Effective Visualization and Trends Forecasting of Crime Data. *IEEEAccess*, Volume 7, pp. 106111 - 106123.
- Friendly, F., 2019. Jaro–Winkler Distance Improvement For Approximate String Search Using Indexing Data For Multiuser Application. *Journal of Physics: Conference Series*, Volume 1361, pp. 1-6.

- Fu, X., Boongoen, T. & Qiang, S., 2010. Evidence Directed Generation of Plausible Crime Scenarios with Identity Resolution. *Taylor & Francis*.
- Gan, J. et al., 2023. Underground Garage Patrol Based on Road Marking Recognition by Keras and Tensorflow. *MDPI*, 13(4).
- Goldsborough, P., 2016. *A Tour of TensorFlow*, München: s.n.
- Gordan, G. R., Rebovich, D. J., Choo, K.-S. & Gordon, J. B., 2007. *Identity Fraud Trends and Patterns: Building a Data-Based Foundation for Proactive Enforcement*, s.l.: US Department of Homeland Security United States Secret Service.
- Graves, A. & Schmidhuber, J., 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, p. 602–610.
- Haghighi, S., Jasemi, M., Hessabi, S. & Zolanvari, A., 2018. PyCM: Multiclass confusion matrix library in Python. *The Journal of Open Source Software*, 3(25), pp. 729-730.
- Hastomo, W. et al., 2021. Characteristic Parameters of Epoch Deep Learning to Predict Covid-19 Data in Indonesia. *Journal of Physics: Conference Series*, pp. 1-4.
- Huang, J. & Ling, C., 2005. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering*, 17(3), pp. 299-310.
- Isard, M., Murray, D. G. & Abadi, M., 2017. *A Computational Model for TensorFlow*. Barcelona, ACM, pp. 1-7.
- Jesus, M., 2011. Machine Learning Forensics for Law Enforcement, Security, and Intelligence. In: *Machine Learning Forensics for Law Enforcement, Security, and Intelligence*. Florida: CRC Press, pp. 1-32.
- Kazemian, H., Amirhosseini, M.-H. & Phillips, M., 2022. *Application of Graph-Based Technique to Identity Resolution*. Crete, 18th International Conference on Artificial Intelligence Applications and Innovations.
- Kazemian, H. & Shrestha, S., 2023. Comparisons of machine learning techniques for detecting fraudulent criminal identities. *Expert Systems with Applications*, 229(<https://doi.org/10.1016/j.eswa.2023.120591>).
- Krishnan, S. & Magalingam, P., 2021. *Hybrid deep learning model using recurrent neural network and gated recurrent unit for heart disease prediction*. Kuala Lumpur, International Journal of Electrical and Computer Engineering.
- Kumar, A. et al., 2019. Sarcasm Detection Using Multi-Head Attention Based Bidirectional LSTM. *IEEE Access*, Volume 8, pp. 6388-6397.
- Li, M., Sattar, Y., Thrampoulidis & Oymak, S., 2020. Exploring Weight Importance and Hessian Bias in Model Pruning. *Machine Learning*, 19 June, pp. 1-6.
- Metsker, O., Trofimov, E., Petrov, M. & Butakov, N., 2019. *Russian Court Decisions Data Analysis Using Distributed Computing and Machine Learning to Improve Lawmaking and Law Enforcement*. Saint Petersburg, 8th International Young Scientist Conference on Computational Science.
- Netowl, 2025. *Multiple Uses of Identity Resolution in a Fraud Risk Management System*. [Online] Available at: <https://www.netowl.com/identity-resolution-fraud-risk-management-system> [Accessed 14 February 2025].
- Parisi, L., Ma, R., RaviChandran, N. & Lanzillotta, M., 2021. hyper-sinh: An accurate and reliable function from shallow to deep learning in TensorFlow and Keras. *Elsevier*, Volume 6, pp. 1-5.
- Pastaltzidis, I. et al., 2022. Data augmentation for fairness-aware machine learning. *ACM Digital Library*, pp. 2302-2312.
- Phillips, M., Amirhosseini, M. H. & Kazemian, H. B., 2020. *A Rule and Graph-Based Approach for Targeted Identity Resolution on Policing Data*. Canberra, IEEE.
- Pundhir, S., Kumari, V. & Ghose, U., 2022. Performance Interpretation of Supervised Artificial Neural Network Highlighting Role of Weight and Bias for Link Prediction. *SpringerLink*, Volume 836.
- Rampasek, L. & Goldenberg, A., 2016. TensorFlow: Biology's Gateway to Deep Learning?. *What are its applications for computational biology?*, pp. 12-14.
- Restelli, D., 2017. *Deep Feature Extraction for Sample-Efficient Reinforcement Learning*, Milano: s.n.
- Saravanan, P. et al., 2020. Survey on Crime Analysis and Prediction Using Data Mining and Machine Learning Techniques. In: *Advances in Smart Grid Technology*. Singapore: Springer.
- Sergeec, A. & Balso, M. D., 2018. *Horovod: fast and easy distributed deep learning in TensorFlow*, s.l.: Uber.

Snachez, S., Romero, H. & Morales, A., 2019. *A review: Comparison of performance metrics of pretrained models for object detection using the TensorFlow framework*. Sincelejo, IOP Conference Series: Materials Science and Engineering.

Tardi, C., 2020. What Is the 80-20 Rule?. *Dotdash Portfolio Construction*, 25 May.

Tato, A. & Nkambou, R., 2018. *Improving Adam Optimizer*. Vancouver, International Conference on Learning Representations.

Ullah, I. & Mahmoud, Q. H., 2022. Design and Development of RNN Anomaly Detection Model for IoT Networks. *IEEE Access*, Volume 10, pp. 62722-62750.

United Nations, 2011. Economic Fraud and Identity-related Crime.

Zaccone, G., 2016. Get up and running with the latest numerical computing library by Google and dive deeper into your data !. In: *Getting Started with TensorFlow*. 1st Edition ed. Birmingham, Mumbai: Packt Publishing Ltd., pp. 7-16.

Zhao, Y. & Teng, C., 2025. Classification of soil layers in Deep Cement Mixing using optimized random forest integrated with AB-SMOTE for imbalance data. *Computers and Geotechnics*, Volume 179.