CRIME PREDICTION AND ANALYSIS USING BIG DATA

[1,1]Anjana Ravi, [2,1]Praseetha V.M

[1]M.Tech Student, [2]Associate Professor
[1]Department of Computer Science and Engineering, St. Joseph's College of Engineering and Technology, Pala, Kottayam, APJ Abdul Kalam Technological University, Kerala, India.

Abstract: Big data analysis is an approach for identifying and analyzing patterns, trends and the unknown relations between the data. It is often a complex process of analyzing a large volume of various data. The use of deep learning techniques in big data analytics is to optimize the data and to analyze the data more precisely. The data visualization shows different facts about the crime data. In this paper, we construct two deep learning models, LSTM and Stacked LSTM. And we predict the crime type using these models. The models are compared using the accuracy performance of crime prediction. The result shows that the Stacked LSTM is better in prediction accuracy than the LSTM. And this comparison can be used for finding a better model. The experimental results using these models can be used for making decisions in police departments and law enforcement organizations.

Index Terms **- Deep learning, LSTM, Stacked LSTM, Data visualization, Time series forecasting.**

I. INTRODUCTION

Big data analytics is highly emerging in the recent years for analyzing different data and extracting information such as the hidden patterns and the unknown correlation between the data [1]. The big data analytics have a wide range of applications and can handle the issues of data that are too vast, too unstructured, and too fast moving to be managed by traditional methods [8]. We can achieve new insights for understanding the patterns of such data.

In big data analytics, deep learning is one of the trending technologies. Deep learning techniques are significantly implemented to numerous fields of technology and engineering along with speech recognition, picture classifications, and learning techniques in language processing. There are different deep learning models. Recurrent Neural Network is an example, in which the data is trained in the networks with feed-back mechanism. But, RNN suffer from the vanishing gradient problem. So, to overcome the vanishing gradient problem and to deals with the long dependencies Long Short Term Memory model is used.

The population expansion and the urbanization led the society to face different violent crimes. The study in crimes will give an awareness of how the crime rates are increasing. So to deal with such problems, we have different techniques [4]. The data related to crimes are increasing exponentially. So, we should find out different methods to analyze the large amount of heterogeneous crime data [5].

In this paper, we have constructed stacked LSTM and single LSTM model for training the crime data. The two deep learning models and visualization techniques are used to process the crime data. We explore the data in Chicago crime data set and we compare how the model prediction accuracy in Long Short Term Memory model and Stacked LSTM. We study about how the number of epochs and batch size affect the prediction accuracy of two the models.

The rest of the paper organized in the following way. Section 2 discusses about the related works and Section 3 describes the proposed system, summarizes the data set analysis, data visualization and describes the deep learning models for crime prediction. Experimental process and result analysis are presented in Section 4 and Section 5, respectively, followed by some conclusions.

II. RELATED WORK

In recent years, there is a drastic growth in crimes and thereby the crime rates are increased. The increasing population growth is one of the major reasons for the increase in crime rates [1]. The crime types are predicted and after prediction the crime analysis helps to detect the area which has the greatest number of crimes. The area which is having the highest number of crimes is called 'Hotspots'[3]. By analyzing the crime using different approaches can find out different possibilities of occurring crime offences. The analysis of crime will help us to know how the crime trends are going. The crime analysis can be done in several steps, from classification to visualizing the results [2].

In [4] the authors discuss about the time series forecasting or time series prediction in crime data. In time series forecasting, the prediction is based on past events, that is, time-series data. The set of observations on the values that a variable takes at different times is defined as the time series data. Here, they are also discuss about the comparison of performance evaluation between the deep learning models using crime data.

Neda *et al.*[18] discuss about the ARIMA model and LSTM model. ARIMA Model is considered as the traditional model and the model only deals with the stationary data but the LSTM is the deep learning model and can take different kinds of data. The comparison between the ARIMA model and LSTM model are also discussing. In [6], the authors discuss the comparison between

the forecast effects of the LSTM and ARIMA model. The ARIMA model is an integrated model by integrating Autoregressive model that is, AR model and Moving Average model that is, MA model.

In [13] the authors say that the crime data are increasing. For analyzing the crime data there are different methods and algorithms. Crime analysis depends on different attributes and major factor is the type of crime. In [7] and [19] authors discuss about the steps involves in the crime analysis and prediction. Yadav *et al.*[8] discuss about the ARIMA model. The ARIMA Model works well for managing the data and the model can reduce prediction error. The crime data set used for crime analysis is the data taken from the National Crime Record Bureau (NCRB). Zheng *et al.*in their paper [1] discuss about crime analysis using neural networks. The model is trained using the three publically available datasets which are, San-Francisco, Chicago, and Philadelphia crime data sets. In [13] the LSTM model is trained using the Chicago data set and Los Angeles data set. The dataset of Chicago and Los Angeles cities contain the historical information of crimes from 2001 to 2019 and 2010 to 2018 respectively.

III. PROPOSED SYSTEM

In our proposed system, we conducted data analysis, data visualization and crime prediction using LSTM and stacked LSTM model and compared the results. First we train the crime data with stacked LSTM and single LSTM. Then we compare the performance of stacked LSTM and single LSTM with the accuracy of crime prediction. Data visualization techniques are used to extract the relationship between the different attributes in the crime data set. Deep learning models are used for crime prediction and analysis to find the optimal model. Fig 3.1 shows the proposed system architecture. We used the Chicago crime data set for data exploration and crime prediction and detailed description of data set is in next section. Fig 3.2 shows the workflow of proposed system.

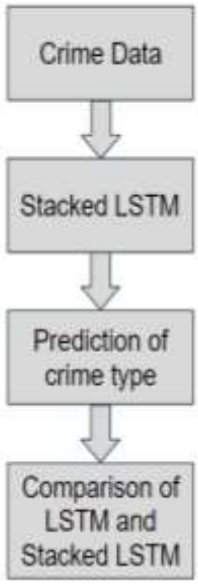


Fig 3.1 Proposed System Architecture

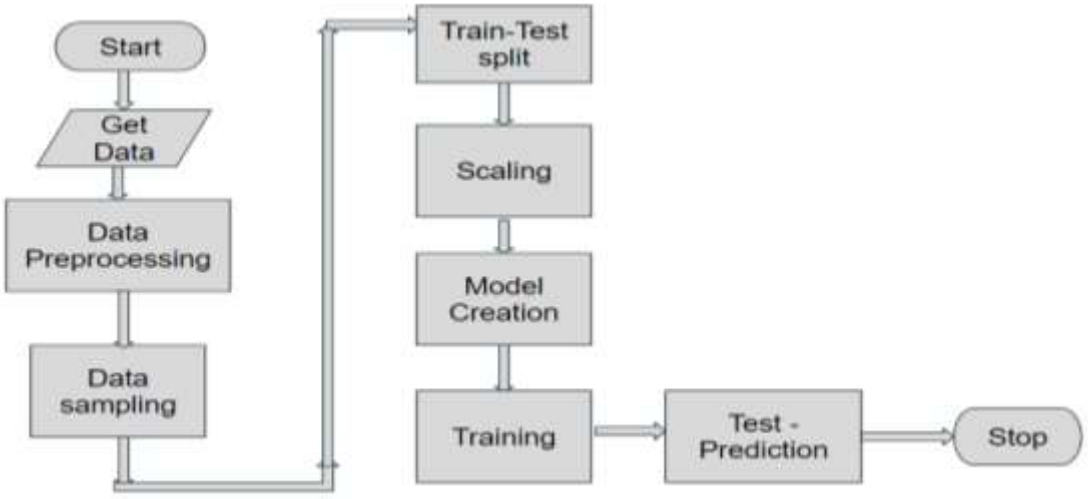


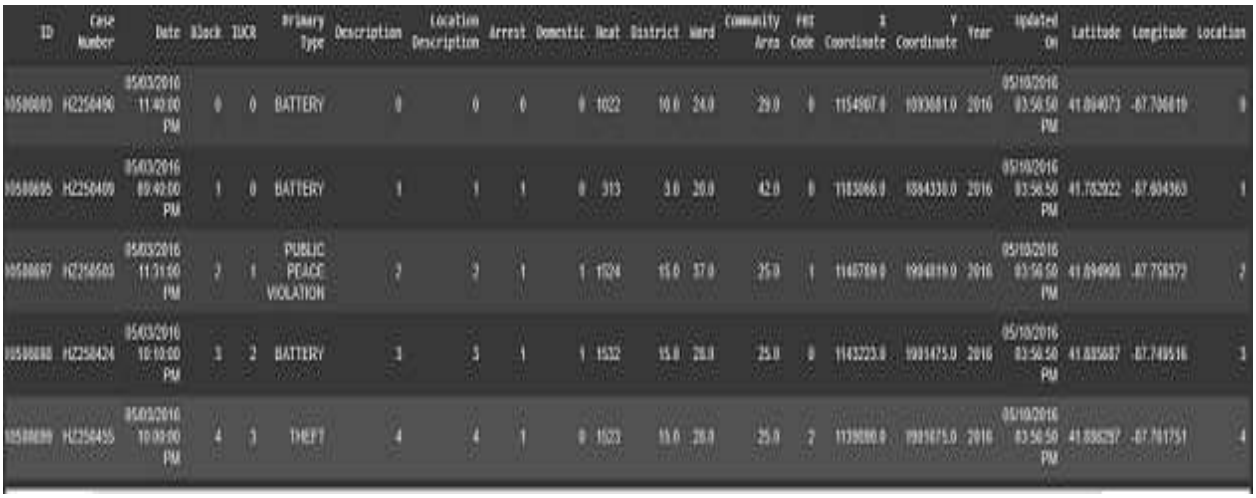
Fig 3.2 Work flow of Proposed System

3.1 Dataset Analysis

For this experiment we are using Chicago crime Dataset and this dataset is available in Kaggle dataset source. Fig 3.3 shows the Chicago crime data set.

The Chicago crime data contains the crime incidents from 2012 to 2017 and it includes,

1) ID – Unique Identifier for the record.
2) Case Number – The Chicago Police Department Number which is unique to the incident.
3) Date – Date when the incident occurred.
4) Block – The partially address where the incident occurred.
5) IUCR – The Illinois Uniform Crime Repository code which is related to primary type.
6) Primary Type – The type of the crime.
7) Description - The secondary description of the IUCR code, a subcategory of the primary description.
8) Location Description - Description of the location where the incident occurred.
9) Arrest - Indicates whether an arrest was made.
10) Domestic - Indicates whether the incident was domestic-related as defined by the Illinois Domestic Violence Act.
11) Beat - Indicates the beat where the incident occurred.
12) District - Indicates the police district where the incident occurred.
13) Ward - The ward (City Council district) where the incident occurred.
14) Community Area - Indicates the community area where the incident occurred. Chicago has 77 community areas.
15) FBI Code - Indicates the crime classification as outlined in the FBI's National Incident-Based Reporting System (NIBRS).
16) X Coordinate - The x coordinate of the location.
17) Y Coordinate - The y coordinate of the location where the incident occurred.
18) Year - Year the incident occurred.
19) Updated On - Date and time the record was last updated.
20) Latitude - The latitude of the location where the incident occurred.
21) Longitude - The longitude of the location where the incident occurred.
22) Location - The location where the incident occurred.



ID	Case Number	Date	Block	IUCR	Primary Type	Description	Location Description	Arrest	Domestic	Beat	District	Ward	Community Area	FBI Code	X Coordinate	Y Coordinate	Year	Updated On	Latitude	Longitude	Location
10508693	HZ250496	05/03/2016 11:40:00 PM	0	0	BATTERY	0	0	0	0	1022	10.0	24.0	29.0	0	1154907.0	1893681.0	2016	05/10/2016 03:56:50 PM	41.864073	-87.706819	0
10508695	HZ250409	05/03/2016 09:40:00 PM	1	0	BATTERY	1	1	1	0	315	3.0	20.0	42.0	0	1183066.0	1884330.0	2016	05/10/2016 03:56:50 PM	41.782922	-87.604363	1
10508697	HZ250503	05/03/2016 11:31:00 PM	2	1	PUBLIC PEACE VIOLATION	2	2	1	1	1524	15.0	37.0	25.0	1	1140789.0	1894819.0	2016	05/10/2016 03:56:50 PM	41.894998	-87.758372	2
10508888	HZ250424	05/03/2016 10:10:00 PM	3	2	BATTERY	3	3	1	1	1532	15.0	28.0	25.0	0	1143223.0	1889475.0	2016	05/10/2016 03:56:50 PM	41.880687	-87.749516	3
10508899	HZ250455	05/03/2016 10:00:00 PM	4	3	THEFT	4	4	1	0	1523	15.0	28.0	25.0	2	1139080.0	1891675.0	2016	05/10/2016 03:56:50 PM	41.886297	-87.765751	4

Fig 3.3 Chicago crime data set

3.2 Data Preprocessing

Data preprocessing is the process which change the data for further processing and training. The data preprocessing includes data cleaning, handling missing values etc. We check whether the data contain any null values or any data is missing in the data set. Then we use dropna() method to drop all the null values in the data set. After the preprocessing step the data is split into train data and test data. 75 % of total data is used for training and 25% is used for testing. The model is training using the train data and we test the model with the test data.

Major steps in preprocessing:
1) Drop all the null values.
2) Drop the unwanted columns.
3) Split the date attribute in the dataset into month, week, day, hour.
4) Change the character field to numeric data.
5) Scaling of both train data and test data.
6) Train-Test-Split.

3.3. Data Visualization

Data visualization gives an idea of how the data in the data set are distributed. We can plot different graphs for finding the trends in the data set. Here, we use the Tableau software for plotting the graphical representation.
The Chicago data set visualization can be done for monthly occurrence of crime, crimes happened in hour, crime trends etc. Fig 3.4 shows the monthly crime trends in Chicago crime data from 2012 to 2017. In this we can see that from January to March there is very less crime compared to other months and also the crime rate is low in the month December also. The monthly analysis of crimes shows May is the month which is having more crime occurrences and February is the month with less crime occurrences. Fig 3.5 summarizes the hourly crime trends and it shows that during the mid of the day the crimes are increasing. In morning hours, the number of crimes is less compared to other hours.
Fig 3.6 shows the number of crime type in Chicago. There are different types of crimes happened in Chicago. From this we can get an idea that theft is the major crime happened in Chicago during 2012 to 2017 and battery, unlawful application of physical force to another person, is the second most crime happened. So, the data visualization gives a clear idea about the trends, patterns and outliers in a large data set. The crime trends are varying in every month and in every hour.

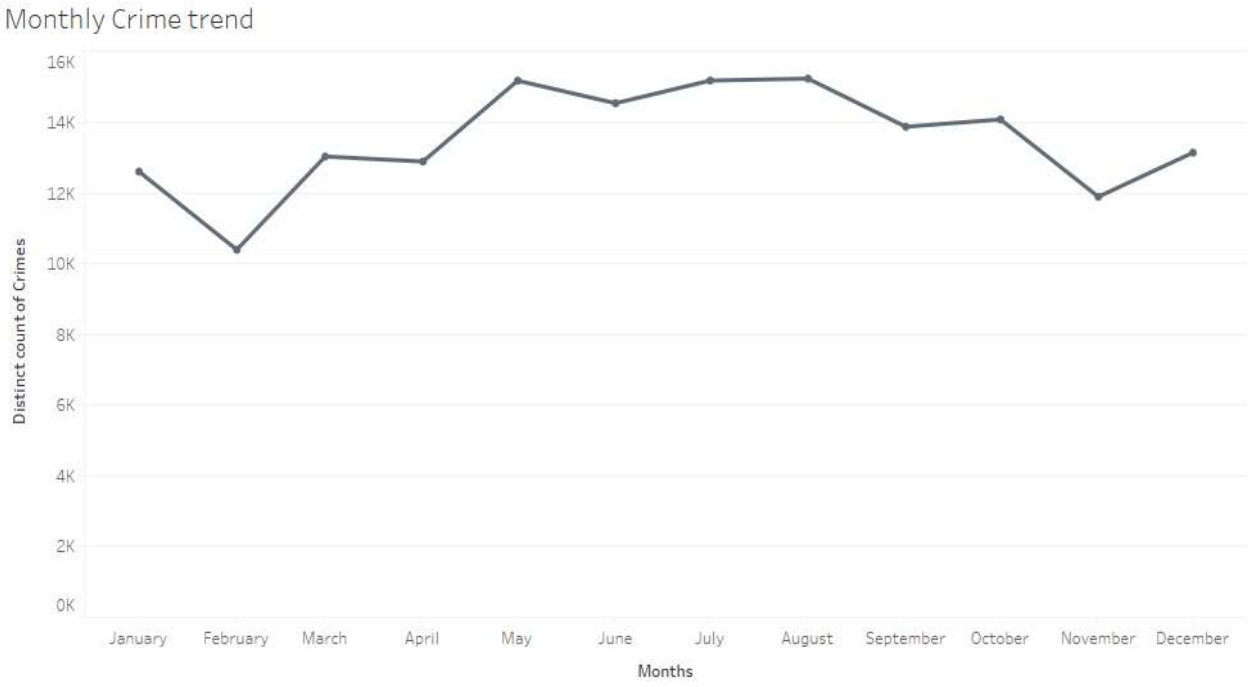


Fig 3.4 Monthly crime trends

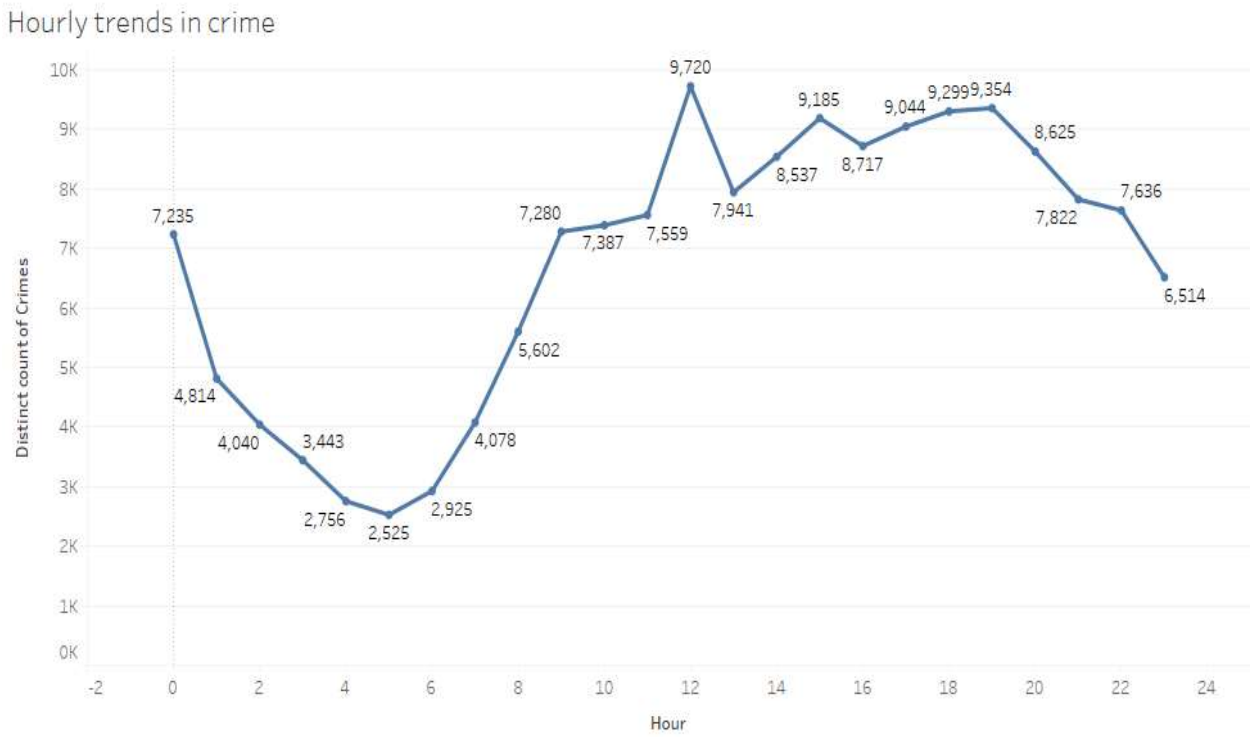


Fig 3.5 Hourly crime trends

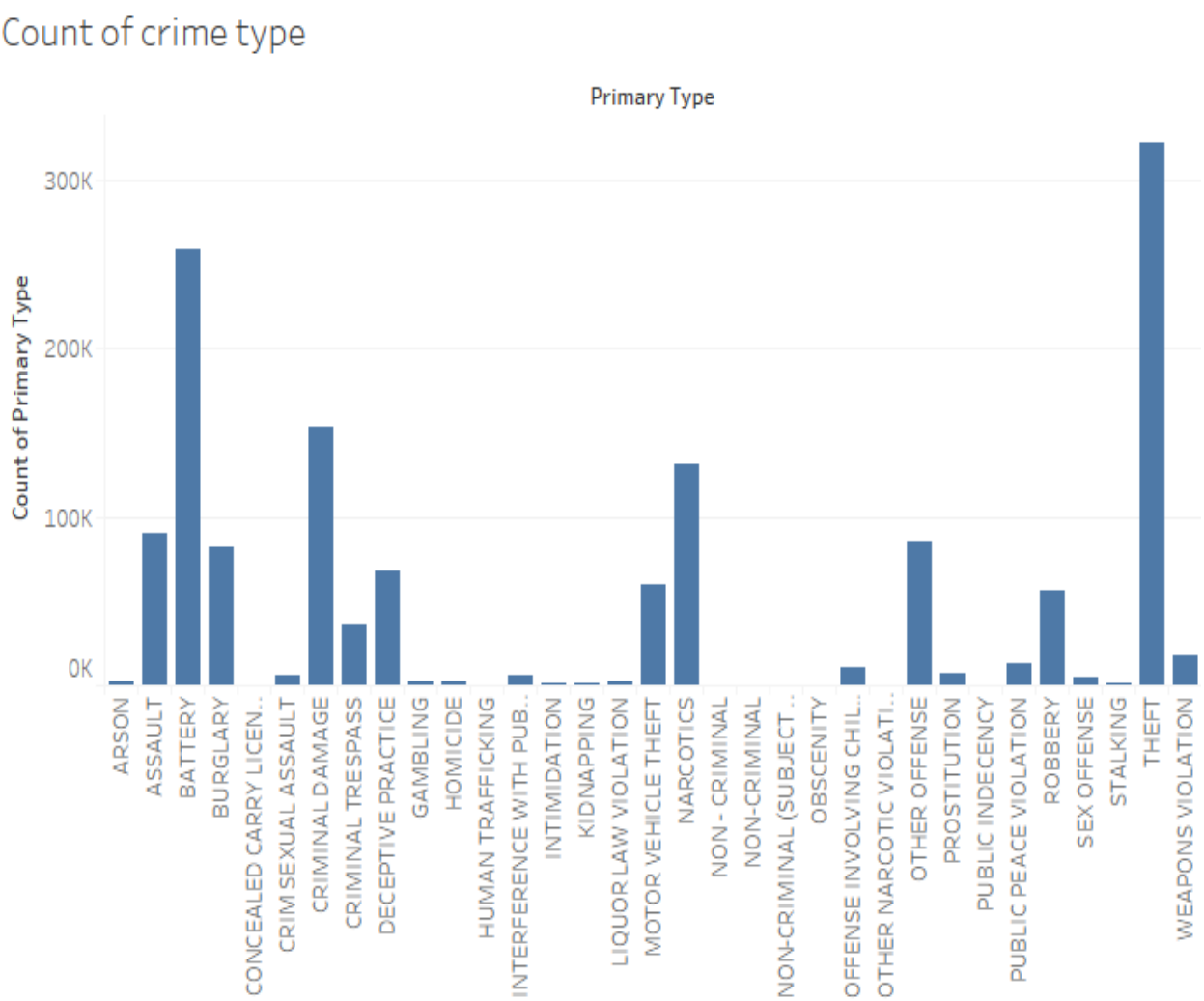


Fig 3.6 Number of Crime type

3.4 Prediction Models

Several machine learning and deep learning models are used to find the crime prediction. The prediction models should be trained with the data to get more accurate results. There are different deep learning models used to train the crime data set. Deep learning models can give better results than the machine learning models. Deep learning models have the feature of deep training with the parameters in the data set. The deep learning models like LSTM have the feature of carrying the parameters for long duration of time and keep the dependencies between the parameters for long time. LSTM models can memorize the dependencies for a long period. So, it is called Long Short Term Memory model.

3.4.1 Crime Prediction using LSTM Model

LSTM model is the type of recurrent neural network which includes the feature of long short term dependencies. In LSTM there are three gates and a cell state. The gated mechanism in LSTM provides the ability to add or remove information to the cell state. Forget gate decides whether the information to be retained or not. The input gate will decide what new information is getting stored in the cell state. In this work, we used the LSTM model for sequence classification problem. Here, the model will predict the trends in crime with the historical data which is trained in the LSTM model. The preprocessing steps and the training of the model will give a better result compared to other models. Fig 3.7 Shows the LSTM model and fig 3.8 shows the summary of the model we created.

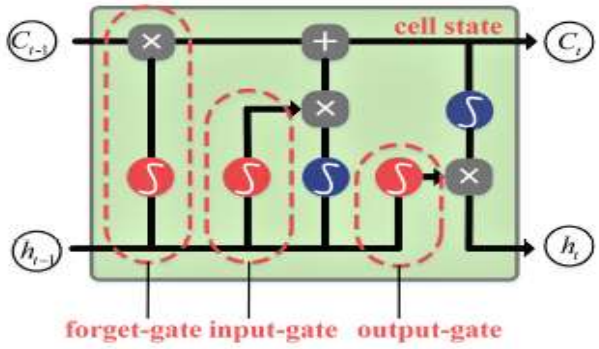


Fig 3.7 LSTM model

```
Model: "sequential_1"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_2 (LSTM)                (None, 1, 50)             13000
_________________________________________________________________
dropout_2 (Dropout)          (None, 1, 50)             0
_________________________________________________________________
lstm_3 (LSTM)                (None, 50)                20200
_________________________________________________________________
dense_1 (Dense)              (None, 20)                1020
_________________________________________________________________
dropout_3 (Dropout)          (None, 20)                0
=================================================================
Total params: 34,220
Trainable params: 34,220
Non-trainable params: 0
```

Fig 3.8 Summary of LSTM model.

3.4.2 Crime Prediction using Stacked LSTM

Stacked LSTM is the model which can be made by stacking the single LSTM model. In this model, each LSTM will get a sequence of output from the other layers rather than a single output value. Stacked LSTM have more layers for training and each layer is connected to other layer in the LSTM. In crime prediction and analysis the stacked LSTM will give better results than the single LSTM model. Stacked LSTM is a complex model but this will train the crime data better than the other models. Fig 3.9 shows the model summary of stacked LSTM that we created and fig 3.10 shows the stacked LSTM model.

```
Model: "sequential_2"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_21 (LSTM)               (None, 14, 50)            10400
_________________________________________________________________
lstm_22 (LSTM)               (None, 14, 50)            20200
_________________________________________________________________
lstm_23 (LSTM)               (None, 50)                20200
_________________________________________________________________
dense_2 (Dense)              (None, 20)                1020
=================================================================
Total params: 51,820
Trainable params: 51,820
Non-trainable params: 0
```

Fig 3.9 Summary of Stacked LSTM model

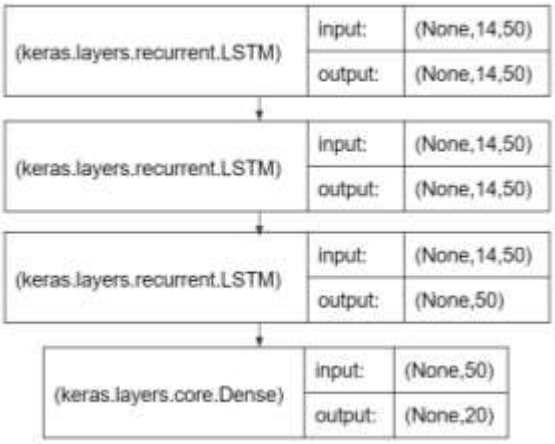


Fig 3.10 Stacked LSTM model

IV. EXPERIMENTAL PROCESS

In this, the LSTM and Stacked LSTM is trained with the Chicago crime data set. After the preprocessing and train-test-split we created single LSTM model and stacked LSTM model with 50 hidden layers. Since it is a classification crime prediction problem we created the models using cross entropy as the loss function and softmax as activation function. Dropout refers to ignoring neurons during the training phase of certain set of neurons which is chosen at random. This will help the model from over fitting and improve the performance of the model. The dropout of the model set to 0.2. Batch size refers to the number of training examples utilized in one iteration or one epoch. The batch size is set to 256. Both the models are trained with different epochs and we got different accuracy for LSTM and stacked LSTM model. We sampled the data to 100000 and in this 75% of sampled data is used as train data and 25% of data is used as test data. Table 4.1 shows the summary of parameters used.

Table 4.1 Parameters used in training

Parameters	Values
Total Data	100000
Train Data	75%
Test Data	25%
Drop out	0.2
Batch Size	256
Hidden layers	50
Activation Function	Softmax
Loss Function	Categorical Cross entropy

V. RESULTS ANALYSIS

We trained the stacked LSTM and single LSTM with the Chicago crime data set. Here we compared the performance of the models with the accuracy in prediction. We trained the models in different epochs and in every epochs we can see the accuracy variation in both the models. Table 5.1 shows the accuracy in different epochs of Stacked LSTM and single LSTM. The accuracy comparison of both prediction models clearly shows that the stacked LSTM model performs better than the single LSTM model when the number of epochs increases. Fig 5.1 shows the accuracy comparison of stacked LSTM and single LSTM. In every epochs the stacked LSTM model gives better accuracy than the single LSTM model. From the table it is clear that the stacked LSTM model gives maximum accuracy when the epoch is 100. The performance of the stacked LSTM model is better than the single LSTM model since the data set is well trained in the stacked LSTM model. The output from every LSTM layer is trained through all the LSTM layers in the stacked LSTM model. So, the accuracy is better than the LSTM model.

Table 5.1 Accuracy comparison of Stacked LSTM and single LSTM

Number of epochs	Stacked LSTM	LSTM
10	0.26	0.41
20	0.69	0.53
30	0.77	0.47
40	0.83	0.54
50	0.85	0.51
60	0.87	0.60
70	0.88	0.60
80	0.80	0.64
90	0.87	0.68
100	0.90	0.73

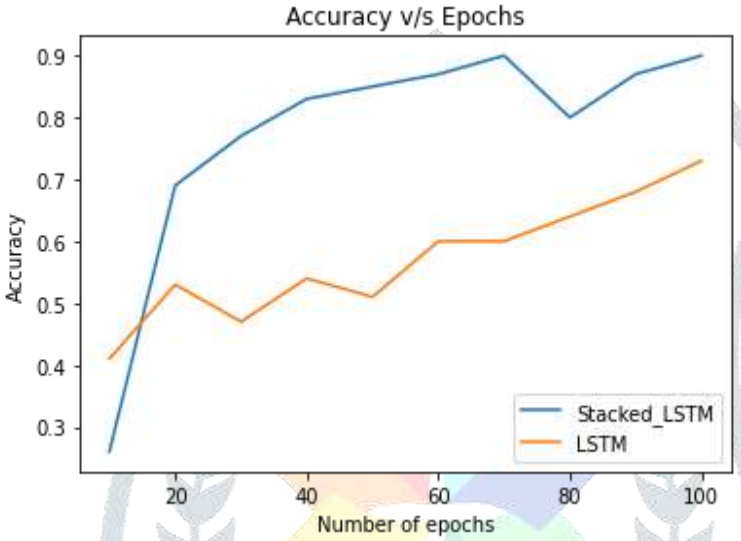


Fig 5.1 Accuracy comparison of Stacked LSTM and single LSTM

We can also compare the loss of stacked LSTM and single LSTM. The comparison of loss parameter shows that the stacked LSTM have less loss during the training process. Table 5.2 clearly shows that when epoch is 10 the stacked LSTM have minimum loss and the single LSTM have maximum loss. The loss of both models is decreasing but the overall performance of stacked LSTM is better than the single LSTM model. So, by comparing the performance of prediction accuracy and loss it can be seen that stacked LSTM works better than the single LSTM model. Fig 5.2 shows the comparison of loss in stacked LSTM and single LSTM.

Table 5.2 Loss comparison of stacked LSTM and LSTM

Number of epochs	Stacked LSTM	LSTM
10	0.04	2.00
20	0.96	1.60
30	0.67	1.50
40	0.49	1.40
50	0.46	1.40
60	0.37	1.10
70	0.31	1.10
80	0.01	1.06
90	0.31	0.90
100	0.27	0.80

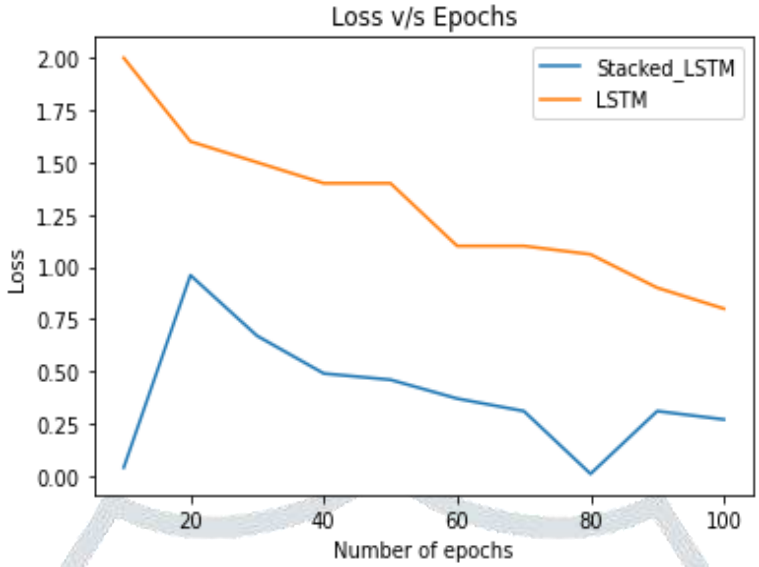


Fig 5.2 Loss comparison of stacked LSTM and LSTM

CONCLUSION & FUTURE WORK

In this paper, we used two deep learning models to analyze the performance in crime prediction of Chicago crime data set. There are different deep learning models and the training of the data in these models will give better results than other traditional models. Here, we trained the Chicago data set using single LSTM and stacked LSTM model. We trained the model with different parameters in different number of epochs and the accuracy is varying for each epochs. By analyzing the results of the performance of stacked LSTM and single LSTM we find out that stacked LSTM model will perform better than the single LSTM model. So, this model can be used by both police department and law enforcement department in crime analysis for decision making.

As a future work, the stacked LSTM model can be trained with different parameters of different types of crime data and we can make an ensemble model to reduce the generalization of the crime prediction. We will conduct more realistic study to improve the effectiveness of the crime prediction.

REFERENCES

[1] Mingchen Feng, Jiangbin Zhengi, Jinchang Ren, Amir Hussian, Xiuxiu LI5,Yue Xi and Qiaoyuan Liu “Big Data Analytics and
Mining for Effective Visualization and Trends Forecasting of Crime Data” ,DOI:*10.1109/ACCESS.2019.2930410*

[2] U. Thongsatapornwatana, ``A survey of data mining techniques for analyzing crime patterns," in *Proc. 2nd Asian Conf. Defence Technol.*, Chiang Mai, Thailand, 2016, pp. 123_128.

[3] Umair Muneer Butt, Sukumar Letchmunan, Fadratul Hafinaz Hassan,Mubashir Ali, Anees Baqir, Hafiz Husnain Raza Sherazi “Spatio-Temporal Crime HotSpot Detection and Prediction: A Systematic Literature Review”, IEEE Transactions and Journals, 10.1109/ACCESS.2017.

[4] K. R. S. Vineeth, T. Pradhan, and A. Pandey, ``A novel approach for intelligent crime pattern discovery and prediction," in *Proc. Int. Conf. Adv. Commun. Control Comput. Technol.*, Ramanathapuram, India, 2016, pp. 531_538.

[5] Z. Zhao, S. Tu, J. Shi, and R. Rao, ``Time-weighted LSTM model with redefined labeling for stock trend prediction," in *Proc. IEEE 29th Int. Conf. Tools Artif. Intell. (ICTAI)*, Boston, MA, USA, Nov. 2017, pp. 1210_1217.

[6] J. Dai, G. Sheng, X. Jiang, and H. Song, ``LSTM networks for the trend prediction of gases dissolved in power transformer insulation oil," in *Proc. 12th Int. Conf. Properties Appl. Dielectr. Mater.*, Xi'an, China, 2018, pp. 666_669.

[7] H. Hassani, X. Huang, M. Ghodsi, and E. S. Silva, ``A review of data mining applications in crime," *Stat. Anal. Data Mining, ASA Data Sci. J.*, vol. 9, no. 3, pp. 139_154, Apr. 2016.

[8] Romika Yadav, Savita Kumari Sheoran,” Crime Prediction Using Auto Regression Techniques for Time Series Data”, 3rd International Conference and Workshops on Recent Advances and Innovations in Engineering, 22-25 November 2018, 978-1-5386-4525-3/18.

[9] I. N. da Silva and D. H. Spatti, ``Introduction in Artificial Neural Networks‘. Cham, Switzerland: Springer, 2017, pp. 3_19.

[10] K. Greff, R. K. Srivastava, J. Koutnìk, B. R. Steunebrink, and J. Schmidhuber, ``LSTM: A search space odyssey," *IEEE Trans. Neural Network. Learn. Syst.*, vol. 28, no. 10, pp. 2222_2232, Oct. 2017.

[11] Wei Zhong, Ning Yu, and Chunyu Ai, "Applying Big Data Based Deep Learning System to Intrusion Detection", Big data mining and analytics ,ISSN 2096-0654 03/06 pp181–195 Volume 3, Number 3, September,2020,10.26599/BDMA.2020.9020003

[12] Fang Wang, Menggang Li, Yiduo Mei, Wenrui Li, "Time Series Data Mining: A Case Study with Big Data Analytics Approach", *IEEE Access 10.1109/ACCESS.2017.*

[13] Wajiha Safat, Sohail Asghar, Saira Andleeb Gillani, "Empirical Analysis for Crime Prediction and Forecasting Using Machine Learning and Deep Learning Techniques", *IEEE Access*, Digital Object Identifier 10.1109/ACCESS.2021.3078117

[14] Tahani Almanie, Rsha Mirza and Elizabeth Lor, "Crime prediction based on crime types and using spatial and temporal criminal hotspots", International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.4, July 2015 DOI :10.5121/ijdkp.2015.5401.

[15] K.B.S. Al-Janabi, "A Proposed Framework for Analyzing Crime Data Set using Decision Tree and Simple K-Means Mining Algorithm," in Journal of Kufa for Mathematics and Computer, Vol. 1, No. 3, 2011, pp. 8-24.

[16] Shiju Sathyadevan, Devan M.S" Crime analysis and prediction using data mining" 2014 First International Conference on Networks & Soft Computing(ICNSC2014),10.1109/CNSC.2014.6906719.

[17] Gaurav Yadav, Richa Vasuja "Analysis of Time Series Prediction using Recurrent Neural Networks",*International Journal of Computer Applications (0975 – 8887) Volume 182 – No. 48, April 2019.*

[18] Sima Siami-Namini, Neda Tavakoli, Akbar Siami Namin, "A Comparison of ARIMA and LSTM in Forecasting Time Series" 2018 17th IEEE International Conference on Machine Learning and Applications, 10.1109/ICMLA.2018.00227.

[19] K.R Sai Vineeth , Ayush Pandey, Tribikram Pradhan, "A Novel Approach for Intelligent Crime Pattern Discovery and Prediction", 2016 International Conference on Advanced Communication Control and Computing Technologies (ICACCCT) ISBN No. .978-1-4673-9545-8.

[20] Xu Zhang, Lin Liu, Luzi Xiao, Jiakai Ji, "Comparison of Machine Learning Algorithms for Predicting Crime Hotspots", Volume 8, 2020, *IEEE Access, 10.1109/ACCESS.2020.3028420.*