Detection of suspicious activity using mobile sensor dataand Modified Sub-space K-NN for criminal investigations

Sukhada Aloni*[a], Divya Shekhawat[a]

[a]PaciFIc University, Pratap Nagar, Udaipur, 313024, Rajasthan, India

Abstract

With the bulk availability of mobile sensors, the data collected from them mustn't be wasted. Nowadays the creation of black-box software that collects this data is not a very difficult task. It is possible to detect suspicious unlawful events using this black-box data. In this paper, we present a novel way of doing forensic investigation using a modified sub-space K-NN (MSK) algorithm. The MSK algorithm is capable of detecting suspicious activities from mobile sensor data. Using this technique, we could detect any normal activity versus suspicious activity with 99.7 % accuracy. We expect the future researcher to develop on this idea and build a solid digital forensic system capable of doing bias-free decisions.

Keywords: Forensic, Mobile sensor data, Black box, mobile data collection

1. Introduction

There are at least 20 % raise in forensic criminal cases after 2021, after the COVID-19 outbreak. Most people started working from home and activities inside the house have raised a lot. Due to these multiple fights within the house has raised. There are some cases, where dead bodies were found inside closed doors, but the reason remains unknown. It could be the natural causeof death, such as heart attack, brain hemorrhage, cardiac arrest, etc. or It could be suicide or even it could be murder. Investigation of such cases is very difficult for the policeman. The body postmortem report can reveal many ambiguities. But, for to be sure, investigators can make use of a devicethat is loaded with sensors (i.e., mobile phone).

Our approach in the proposed method that, will give a stepping stone for mobile phone-based digital forensics. There could be other cases other than murder, such as robbery, rape, accidents, etc. where the digital forensics using mobile phone can be very useful.

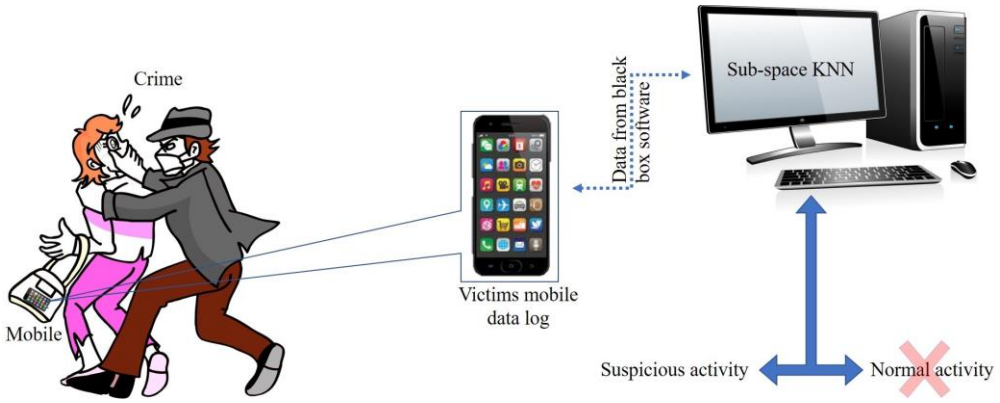

Figure 1: Concept diagram of proposed suspicious activity detector.

2. Literature review

Ferreira *et al.* [1] have designed a survey report on variable electrochemical sensors for different forensic applications. Meffert *et al.* [2] have designed a forensic state acquisition using the Internet of Things. They used an IP camera and intsteion hub for controlling the thermostat. They did cloud testing and found different scenarios that could lead to forensic analysis. They focused on door lock and motion sensor activity. Their entire focus was to design a framework that works with IoT devices. As compared to their work we are avoiding a separate IoT sensor, instead, we use the mobile phone as an IoT device.

Mylonas *et al.* [3] has summarized different smartphone-based forensic techniques and also explained how smartphone-based evidence can be used and their legal implications. Lekshmi *et al.* [4] have designed an SVM-based algorithm that can classify and help in detecting source cameras. They used EXIS information along with sensor pattern noise to determine the exact camera through which the images were acquired. They tried seven different cameras and also considered photo response non-uniformity noise. Based on their work we are also considering using a temperature sensor as one of the

feature parameters. Since it used a unique noise pattern information as perthe environment.

Malik *et al.* [5] provided a network behavior examination method that can allow verification of mobile device forensics. They could not segregate mobile OS-based forensic artifacts however for the known OS they can find behavior of certain devices like ICMP packet transmission or streaming video reception. Their method can determine whether the mobile data is actively generated by the user or passively generated by the device. Jahangiri *et al.* [6] have used mobile senor data for identification of the transportation mode using KNN, SVM, and tree-based classifiers. Similar to their ideas we also going to collect the mobile sensor data but our aim goal is to determine the suspicious activities.

Rosser *et al.* [7] has used mobile phone sensor data for building the interiors and generating the 3-D models. Their approach is to use smartphone sensors to interactively capture the source scene. Their optimizer allows floor plans to get optimized and finally generates models to build the building interiors. Although their application is not directly related to the suspicious activity, we expect our work can also be used to predict future criminal scenes from the available sensor data.

Khan *et al.* [8] have designed a human identification system that can classify human motions based on mobile phone sensors. They used data like accelerometer, Gyroscope, and magnetometer to classify between normal walking and brisk walking. They could reach an accuracy of 96.5 %. Horwitz *et al.* [9] has designed a system that can predict depression and suicidal ideation among medical interns. For the data collection, they used Fitbit and for the prediction of suicidal ideation based on depression detection. They proved that only Fitbit is not good enough for predicting mood.

Ouguz *et al.* [10] has designed a human identification system via accelerometer data from the mobile sensor. They found with 99 % accuracy it is possible to recognize a human based on the accelerometer. For classification, they used KNN and RNN. They validated the dataset with random 3000 samples and 387 devices. Based on their result validation we can say that it is possible to analyze the forensics using mobile sensor data. Our designed system has not only considered the accelerometer but also considered gyroscope and other mobile sensors.

Pradhan *et al.* [11] designed a classification system using wearable sensor technology. They aimed to create reliable IoT technology that can process the information and classify the data as health data or communication data.

Their failure percentage was about 10 % improved over others which are around 0.02. Their response time was 0.9 sec. but the processing time was about 454 milliseconds. Wampfier *et al.* [12] has designed a system that can predict affecting state using smart phone sensors such as touch and heat map. The heat map is created with keystrokes on the smartphones along with gyroscopic and linear acceleration measurements from IMU. Their accuracy could reach up to 70 %.

3. Methodology

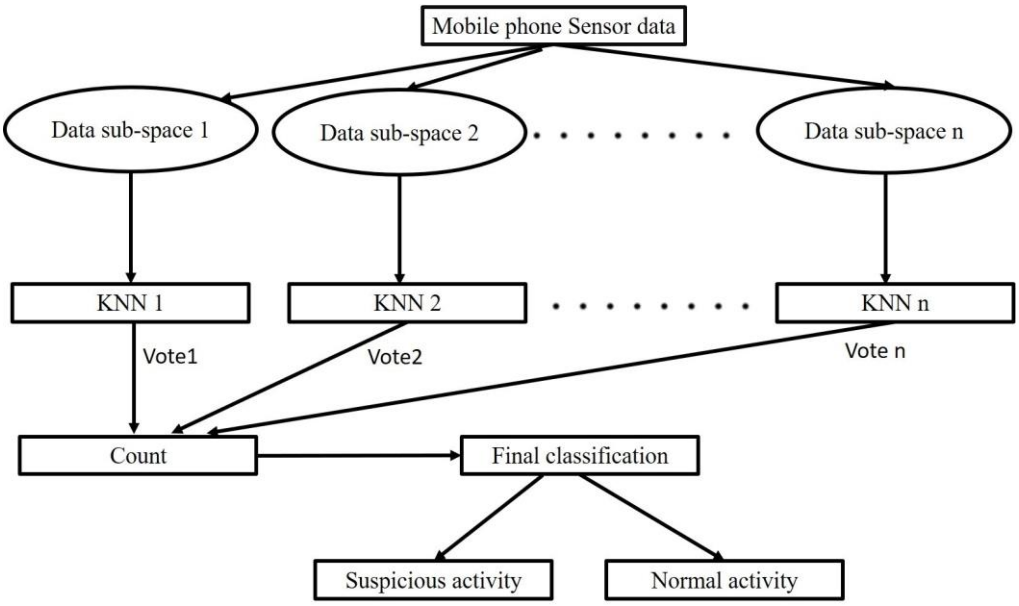

Figure 2: Block diagram for modified sub-space KNN applied for suspicious activity detection from mobile data. The mobile data is collected from a wireless stream and a total of 29 bytes were collected per time instance. Each of the time instance data was subdivided into N subspaces and separate KNN applied. Finally voting from each KNN is counted.The maximum number of votes with their probability is considered for final classification. The suspicious activity is detected if the number of votes is higher than the number of votes to the normal activity.Typically, the number of subspaces was kept odd so that equal votes to both classes can be avoided. In our case, we set this subspace n to 3 and then increase it up to 29 in steps of 2.

To collect the mobile data a black-box application was developed. This application can acquire the GPS coordinates, Accelerometer, Gyroscope,

Magnetic field, Orientation, Linear acceleration, Gravity value, Rotation Value, Pressure, and Battery temperature. The data is collected via the user datagram protocol (UDP) stream. The port on which the communi- cation takes place is 5555. There were 4 different interval selection option (Capture rates). The capture rate was defined as the number pg packets received per unit time. Each packet used in this study consists of 29 bytes of information. The black-box app can run in the background. The app was derived from open source IMU GPS Streaming app. The current study was conducted on Android 10 Funtouch OS_10*andAndroid.*

Fig. 2 shows the block diagram of a modified sub-space KNN used to detect suspicious activity in mobile data. The mobile data was collected from a wireless stream at a rate of 29 bytes per time instance. Each time instance data set was subdivided into N subspaces and each KNN was applied separately. Finally, the votes from each KNN are totaled. For final classification, the maximum number of votes with their probability is taken into account. If the number of votes cast is greater than the number of votes cast for normal activity, suspicious activity is detected. Typically, the number of subspaces was kept odd to avoid giving equal votes to both classes. In our case, we start with a subspace n of 3 and gradually increase it to 29 in steps of 2.

The modified subspace KNN is a method that uses K neighborhood classification with parallel implementation. Here to create, multiple KNN models data is sampled into subspaces, for example out of 999 readings any five readings are randomly chosen and then compared with the current time point.If the number of points is closer to normal activity compared to suspicious activity then the current point is classified as normal. the main modification to the sub-space KNN comes from the fact that we not only give the output class but also give the output probability that the current value belongs to that particular class. For the probability determination, we calculate the sum of the Euclidean distance of the point from each of the K points in the KNN. The formula for euclidian distance is given in equation (1).

$$D_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2} \quad (1)$$

Where,

x_c, y_c are the co-ordinates of current data point being classified.

x_i, y_i are the co-ordinates of i^{th} neighbour, and i varies from 1,2,...K

D_i is the euclidean distance for the i^{th} neigbour and i varies from 1,2,...K.

The formula for probability computation is given in equation 2.

$$P_S = \frac{\frac{1}{L}\sum_{i=1}^{L} D_s i}{\frac{i}{K-L}\sum_{j=1}^{K-L} D_n j + \frac{i=1}{L}\sum_{1}^{L} D_s i} \quad (2)$$

Where,
P_S are probability of suspicious activity.
L Number of neighbours that belongs to suspicious class
$D_S i$ Euclidean distance from current point to all i points that belongs to suspicious class and i varies from 1,2,...L.
$K - L$ Number of neighbours that belongs to normal class
$D_n j$ Euclidean distance from current point to all j points that belongs to normal class and j varies from 1,2,...K-L.

4. Results

Figure 3 shows plot of Error versus capture rate. As the capture rate in packets per second increases, the percentage error in guessing the correct event decreases. As more packets are received, more information is obtained, resulting in greater accuracy. Figure 4 shows plot of capture rate versus time between successive computations. As the capture rate increases the packet received per second increases which allows more data to be processed. As the computing machine is faster and the time between the successive computations is decreasing with the capture rate one can establish the fact thatthe computing machine is in a wait state causing the computations to go low. As more and more packets are received per second time successive com- puting will keep on decreasing till it reaches computational limits. Once the computational limit is achieved the data will be buffered causing the time between successive computations to increase with the capture rate. From the above result, we could successfully reduce the time between back-to-back computations even at the highest capture rate of 200 packets per second. This establishes the speed of mobile and its sensor acquisition could easily be handled by the computing machine used in this experiment. Figure **??**hows feature point plot with the first two principal components marking two different events. The Blue colored dot is for normal activity closed to X= zero,Y= zero and the red-colored dot is the abnormal activities close to X = 3, Y = -30 to +30. This is the PCA plot for one of the twelve different

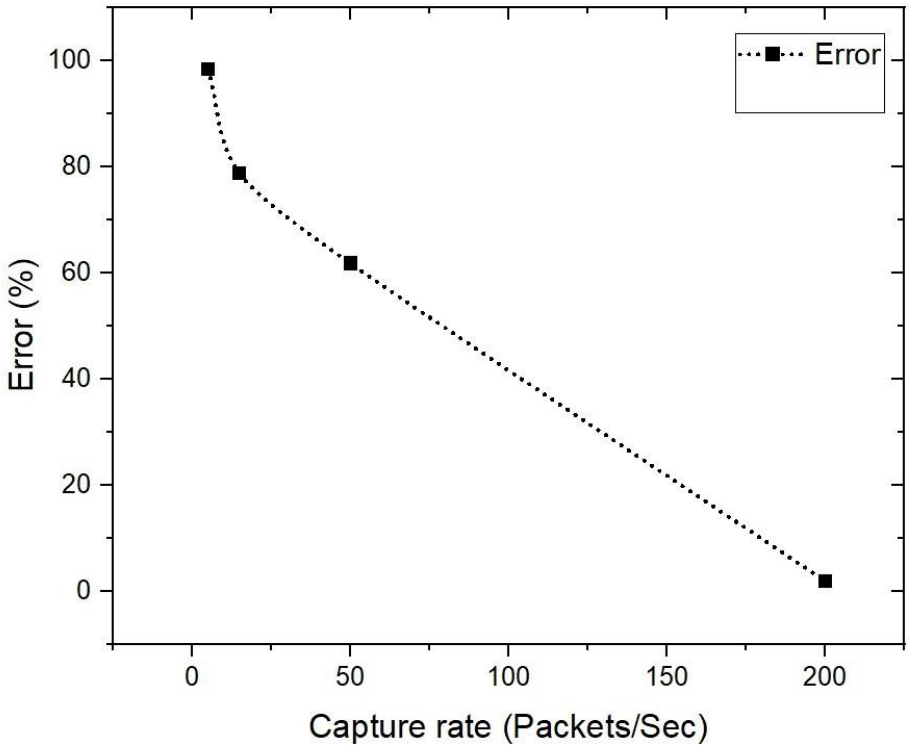

Figure 3: Capture rate versus error. Capture rate increases in packets per second percentage error in guessing the correct event decreases. As more packets are received gives more information and hence results in better accuracy.

sensor readings. Some of the blue dot points from the normal activity is closely mapped with red colored points of abnormal activity. hence only a single sensor feature may not always be sufficient to classify with the highest accuracy of 99.7 % as reported. One should consider all the 12 variables, while getting the final classification plot.

Figure 6 Classifier accuracy confusion matrix 2988 normal events were classified as normal out of 5994 test events, while 9 normal events were classified as suspicious. Out of 2997 suspicious events, 2989 were correctly classified as suspicious, while 8 were incorrectly classified as normal. The modified Sub-space KNN-designed system can classify suspicious events with 99.7 percent accuracy and an average error of 0.28 percent. The ROC curve is represented by a plot of the true positive rate versus the false positive rate.

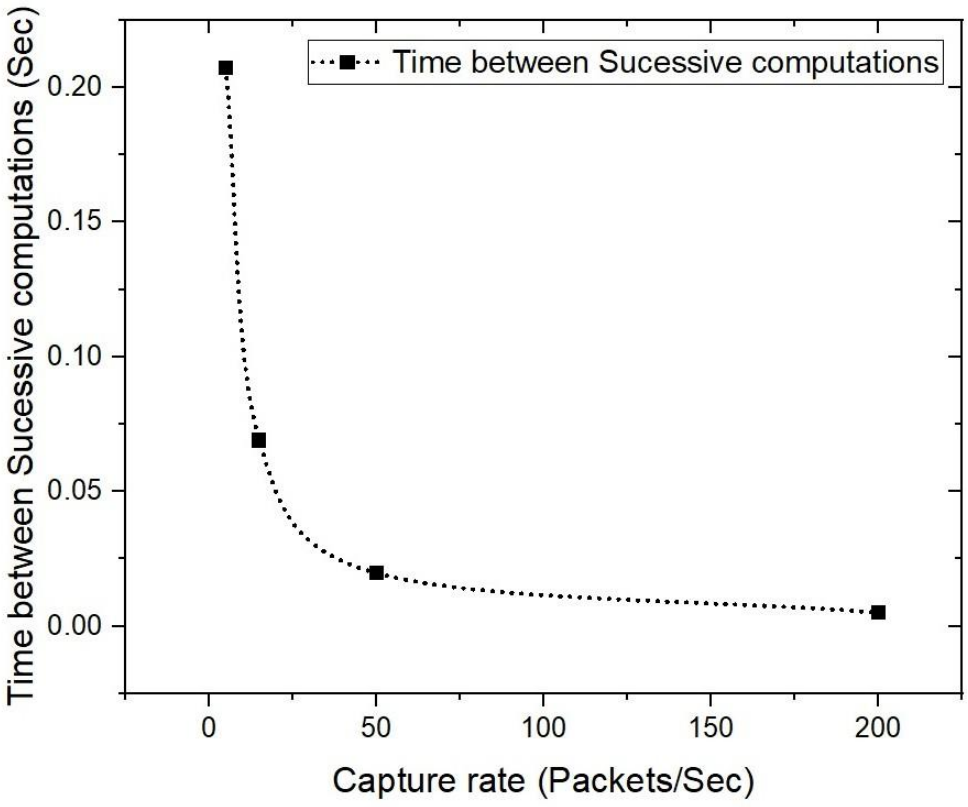

Figure 4: The time between successive computations versus capture rate As the capture rate increases, so does the number of packets received per second, allowing for more data to be processed. As the computing machine becomes faster and the time between successive computations decreases with the capture rate, it is possible to conclude that the computing machine is in a wait state, causing computations to become low. As more packets are received per second, successive computing time decreases until it reaches computational limits.

Auc is the area under the curve. The initial classifier has a true positive rate of exactly one and a false positive rate of zero. There are many ways by which the sensor data could have been collected. We have used a simple app designed in Android studio that requires all the permissions from all the sensors. This particular app works as a black box for human activity monitoring similar to a black box installed in airplanes. Because of this data collection app, we could able to stream the data on a UDP stream and re- ceive the same in another program that was written in Matlab. Although

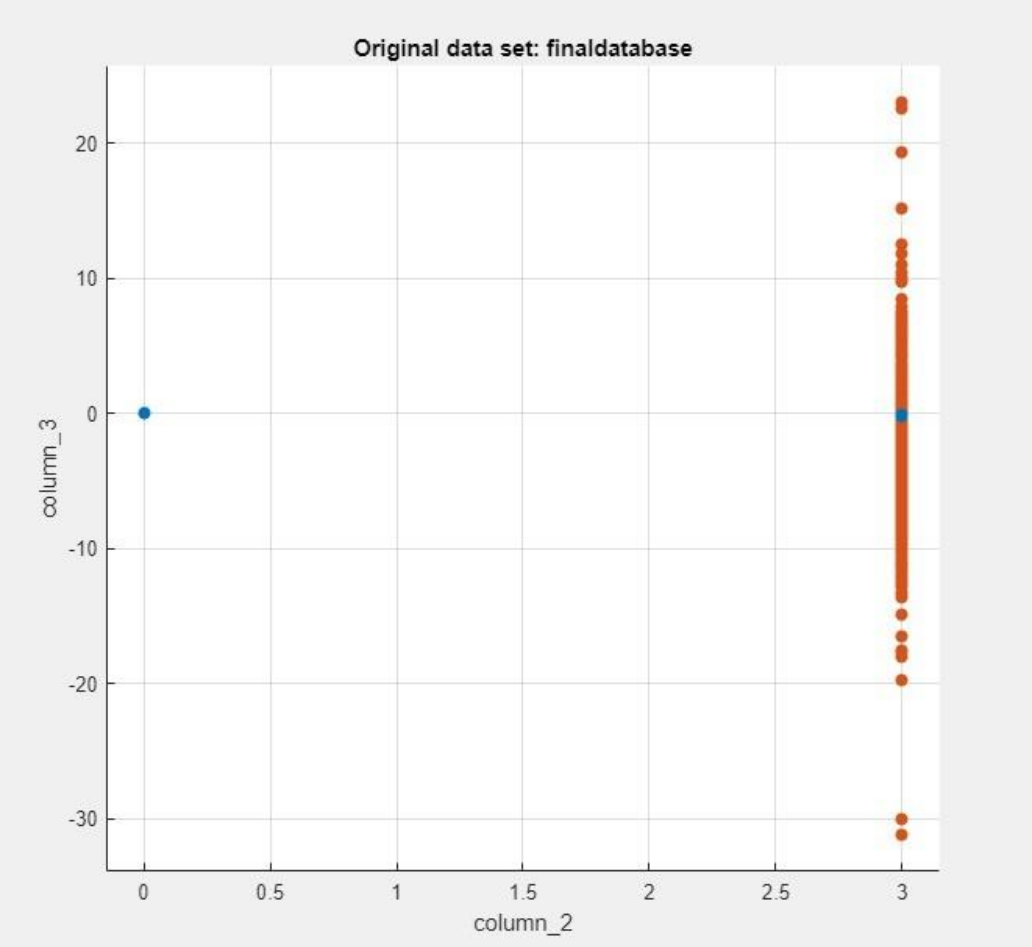

Figure 5: The first two principal components (x-y axis) of the feature point plot represent two distinct events. The blue dot represents normal activity (close to x=0, y=0), and the red dot represents abnormal activity (close to x=3, y=-30 to +30). The PCA plot for one variable of the twelve different sensor reading variables is shown here. Some of the blue dot points representing normal activity are closely mapped with the red colored points representing abnormal activity. As a result, a single variable (sensor) feature may not always be sufficient to classify with the highest reported accuracy of 99.7 percent. When generating the final classification plot one should take into account all 12 variables.

our approach was very straightforward the novelty comes into the picture when we talk about modified Ensemble Sub-space KNN. In this particular approach firstly we divided the entire available data into sub-spaces. These sub-spaces are randomly selected purposefully as it will allow us to follow

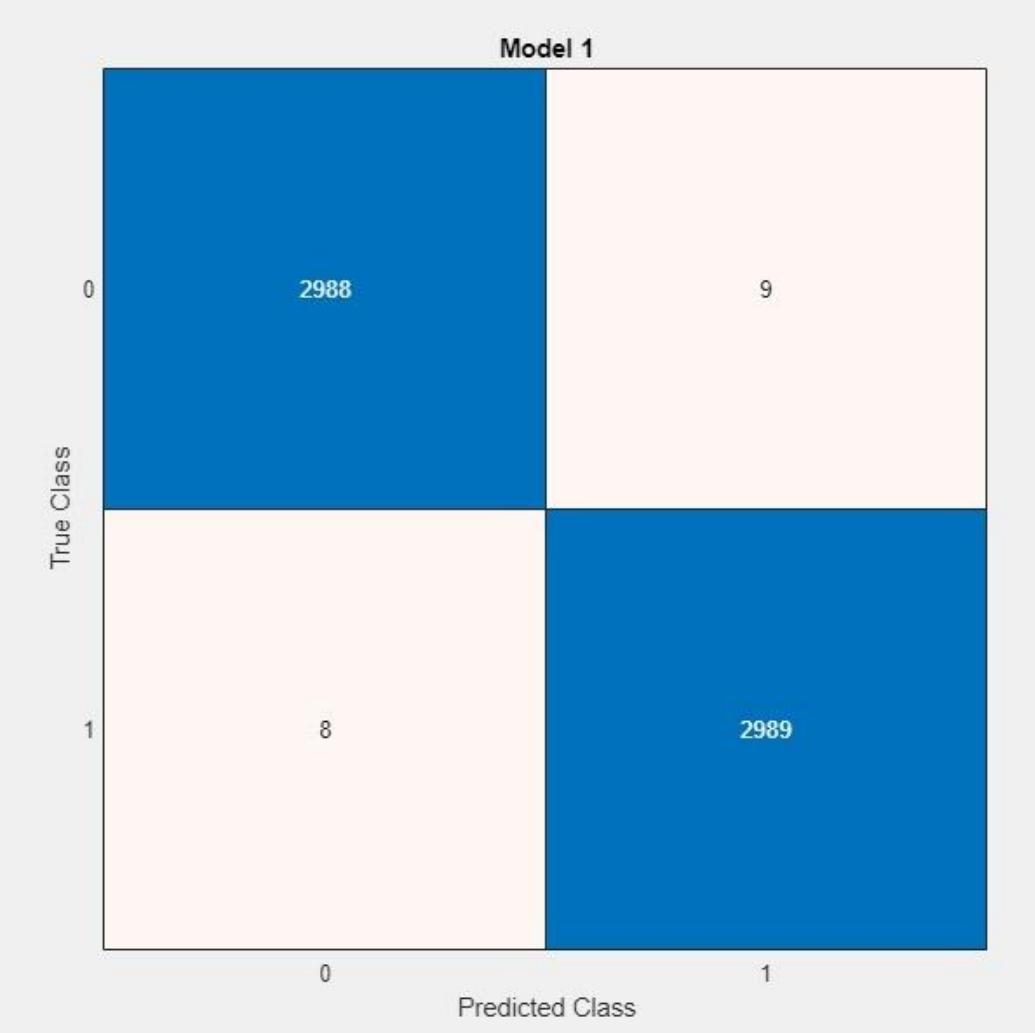

Figure 6: Confusion matrix for classifier accuracy out of 5994 test events 2988 normal events were classified as normal, whereas 9 of the normal events were classified as suspicious events. Out of 2997 suspicious events 2989 were perfectly classified as suspicious events, while 8 such samples were misclassified as normal events. The designed system of modified Sub-space KNN can accurately classify the suspicious event with 99.7 % accuracy and an average error of 0.28 %.

the natural actions which are always random. After sub-spacing the ensembling and maximum voting-based KNN results help us get rid of any biasing present in the output because of the noise. The main modification is the

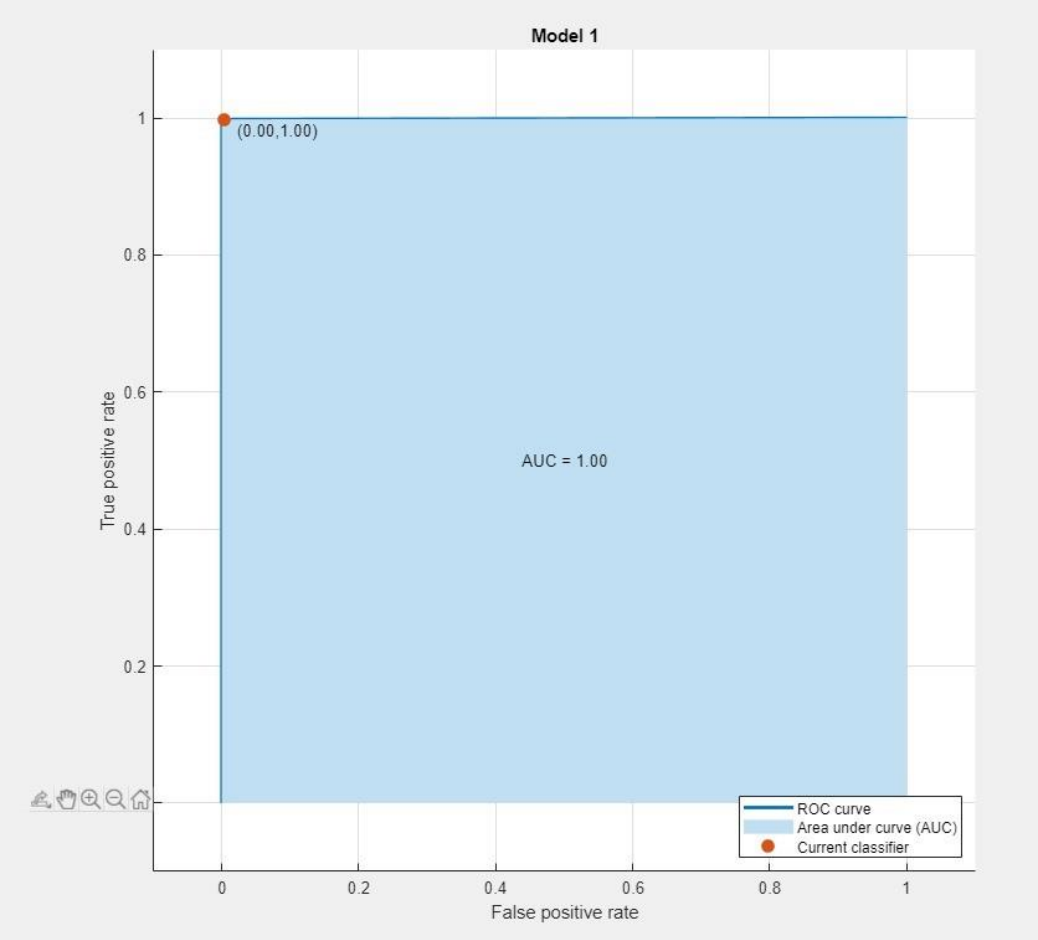

Figure 7: The plot of the True positive rate versus the false positive rate represents the ROC curve. The area under the curve AUC = 1 Initial classifier's true positive rate is exactly 1 at a false positive rate of 0

probability-based KNN where instead of just saying how many neighbors we also say how close these neighbors are with respect to the current point under observation. This allowed us to jump from accuracy of 98 % to 99.7 %. We also tested the classification with the RUSBoosted tree(99.5Because of these reading, we could easily say the proposed algorithm is very efficient with respect to the traditional existing algorithm. We have limited our study to two-class classification but future researchers may explore the opportunity of classifying these mobile data readings into multi-class forensics activity de-

Classification Method	Accuracy (%)
Sub Space KNN	98
RUS Boosted Tree	99.5
Sub space disriminat	83.4
Neural Network	99.4
Modifed Sub space KNN(Proposed algorithm)	99.7

tection. This proposal can be a stepping stone toward digital forensics and we expect these results would improve over time. As the digital data can be updated very easily, we need to provide security structures such that no data editing is possible before and after investigation. By this, we can convince the quotes or judiciary system to accept our proposed mechanism.

5. Conclusions

With the advancements in technology, it is important that humans make most of these available resources. If the sensor data on the mobile phone is stored in form of a black box it will be very useful since it can help investigators to solve the crime. In the proposed work we have also tried to make a black box for a mobile phone that collects all the sensor data and allows the classification of the activity based on this data. For a proof of concept, we have only considered suspicious activity and normal activity as the possible outcome. We achieved an accuracy of 99.7 % which enabled us to justify the proposed digital forensics support system. In this system, we also used a self-developed algorithm called modified subspace KNN. The logic is derived from traditional KNN but the probability factor helps it outperform the existing algorithm. we expect this proposed work will help researchers to build a highly sophisticated digital forensics investigation system. and also modified subspace KNN can be used in various applications.

Author declarations

Funding information

There was no funding involved.

Conflict of Interest

Authors declare that they do not have any conflict of Interest.

Ethics statement

No Animals were involved in the present study. All the data collected from mobile phones of humans who has readily participated in the study and signed a written consent for it. All the necessary permissions are were acquired from institutional ethics committee.

References

[1] P. C. Ferreira, V. N. Ataide, C. L. S. Chagas, L. Angnes, W. K. T. Coltro, T. R. L. C. Paixao, W. R. de Araujo, Wearable electrochemical sensors for forensic and clinical applications, TrAC Trends in Analytical Chemistry 119 (2019) 115622.

[2] C. Meffert, D. Clark, I. Baggili, F. Breitinger, Forensic state acquisition from internet of things (fsaiot) a general framework and practical approach for iot forensics through iot device state acquisition, in: Proceedings of the 12th International Conference on Availability, Reliabilityand Security, 2017, pp. 1–11.

[3] A. Mylonas, V. Meletiadis, B. Tsoumas, L. Mitrou, D. Gritzalis, Smartphone forensics: A proactive investigation scheme for evidence acquisition, in: IFIP International Information Security Conference, Springer, 2012, pp. 249–260.

[4] K. Lekshmi, V. Vaithiyanathan, Source camera identification of image for forensic analysis using sensor fingerprints, in: 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), IEEE, 2018, pp. 1–5.

[5] N. Malik, J. Chandramouli, P. Suresh, K. Fairbanks, L. Watkins, W. H. Robinson, using network traffic to verify mobile device forensic artifacts, in: 2017 14th IEEE Annual Consumer Communications & NetworkingConference (CCNC), IEEE, 2017, pp. 114–119.

[6] A. Jahangiri, H. A. Rakha, Applying machine learning techniques to transportation mode recognition using mobile phone sensor data, IEEE transactions on intelligent transportation systems 16 (5) (2015) 2406–2417.

[7] J. Rosser, J. Morley, G. Smith, Modelling of building interiors with mobile phone sensor data, ISPRS International Journal of Geo-Information 4 (2) (2015) 989–1012.

[8] Y. A. Khan, S. Imaduddin, R. Prabhat, M. Wajid, Classification of human motion activities using mobile phone sensors and deep learning model, in: 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS), Vol. 1, IEEE, 2022, pp. 1381–1386.

[9] A. Horwitz, E. Czyz, N. Al-Dajani, W. Dempsey, Z. Zhao, I. Nahum-Shani, S. Sen, Utilizing daily mood diaries and wearable sensor data to predict depression and suicidal ideation among medical interns, Journal of Affective Disorders (2022).

[10] A. Oğuz, Ö . F. Ertuğrul, Human identification based on accelerometer sensors obtained by mobile phone data, Biomedical Signal Processing and Control 77 (2022) 103847.

[11] M. R. Pradhan, B. Mago, K. Ateeq, A classification-based sensor data processing method for the internet of things assimilated wearable sensor technology, Cluster Computing (2022) 1–16.

[12] R. Wampfler, S. Klingler, B. Solenthaler, V. R. Schinazi, M. Gross, C. Holz, Affective state prediction from smartphone touch and sensor data in the wild, in: CHI Conference on Human Factors in Computing Systems, 2022, pp. 1–14.