Contents lists available at Science-Gate

International Journal of Advanced and Applied Sciences

Journal homepage: http://www.science-gate.com/IJAAS.html

Data mining approach for digital forensics task with deep learning techniques

Lalbihari Barik *

Department of Information Systems, Faculty of Computing and Information Technology in Rabigh, King Abdul Aziz University, Saudi Arabia

ARTICLE INFO

Article history:
Received 21 September 2019
Received in revised form
5 February 2020
Accepted 12 February 2020

Keywords:
Digital forensics
Deep learning
Supervised machine learning
CNN classifiers
Class-based regions

ABSTRACT

In the past, digital forensic, with its exploration techniques, are a lane to the data recovery as well as the examination of different investigation techniques. It is a line of investigation which includes many stages. In this, the foremost assignment is data collection later than that the outcome amount produced predicted with the dataset. Some authors proposed several supervised machine learning techniques that have not obtained much better results. Therefore, the goal of our study was to perform an investigational work on a forensics dataset task for class-based classification methods like three-layer CNN classifiers, five-layer CNN classifiers, and seven-layer CNN classifiers. The classifiers evaluated with classification performance and accuracy. The experimental plan has been done with fivefold cross-validation with fifty repetitions for deep learning algorithms in order to obtain consistent results. Matching accuracy values for the next to next pixels in the classes are calculated with the class-based predicted labels. There are four classes assigned on CNN, and the four classes are segmented and separated with the same region of interest. Then the same class-based region of interests is segregated, and these four class-based regions are next given to CNN with the clusters. Further, the comparison results are made with the used three algorithms.

1. Introduction

Digital forensics investigation has been done with many steps. There are many algorithms included with several stages of data collection, justification, and classifications are processed. Generally, in digital forensics, the primary job is the data collection; it is not an easy task because so many fields will be related to the same investigations (Beebe and Clark, 2005). So, in order to collect the data for real-time, they need to follow or wait for some time. From this, we can conclude that it is a time-consuming process. Digital forensics is a part of digital forensic science in the field of the recuperation, and exploration of stuff originates in digital devices (Conlan et al., 2016). Also, it is frequently used in computer-based crime. Digital forensics is initiated in the establishment of computer forensics, and it has prolonged to envelop the exploration of every device able to store digital data (Reith et al., 2002). Digital forensic investigation has many applications (Shahraki et al., 2013). The most common part of the criminal or civil courts is to carry or disprove a hypothesis before their justification. Criminal cases occupy the suspected violation of the law that is clear by the legislation committee to force by the police and prosecute the cases like theft, murder, and physical attack against the person. Then again, for civil cases for shielding the privileges and belongings of persons, frequently, it is connected amid the family dispute. It is also disturbed through contractual disputes among business-related entities everywhere the digital forensics is very much essential for electronic discovery (Carrier, 2003). Forensics help other than the government sectors such as the private sectors through domestic corporative investigation or infringement investigations. It is also helping the networking area by the specialist investigation criterion into the environment and enlarges the unconstitutional network intrusions.

After the data collection, the output values are predicted from that hidden input data. So,

* Corresponding Author.
Email Address: lbarik@kau.edu.sa
https://doi.org/10.21833/ijaas.2020.05.008
Corresponding author's ORCID profile:
https://orcid.org/0000-0002-5977-6319

researchers proposed more than a few supervised machine learning techniques for these kinds of large dataset forensics work, but they have not obtained to a great extent results. Therefore, in the direction of performance improvement in digital forensics tasks, deep learning techniques are proposed. In this work, the performance of the data mining approach also improved. The technological characteristic investigation separated into some sub-branches. These technological characteristic investigations are completed with the help of digital devices and their types. The digital device forensics are computer and mobile device forensics, network forensics, and forensic data analysis. The classic forensic development includes the capture of forensic image acquisition, digital media analysis, and the construction of a description from the composed proof. In addition to that, the identification of straight crime evidence, digital forensic also helpful for the evidence in exact suspect, confirm alibis or statements, recognize resources, and authenticate documents (Casey, 2009). In the forensic investigation, the searching thing is very small, but the time and human energy are very large.

This paper investigates the natural images investigation work on top of a forensics dataset with the clustering and classification of class-based categorization together with several types of methods. Here there are three-layer CNN classifiers, five-layer CNN classifiers, and seven-layer CNN classifiers are introduced. The classifiers' performances are evaluated based on the classification performance values and accuracy of prediction. The validation process in this work is fivefold cross-validation with fifty repetitions. In deep learning algorithms, the consistent results are obtained only based on the iterations. Matching accuracy standards intended for the next to next pixels in the classes are considered with the class-based prediction labels. Here, four class labels are assigned using CNN layers, and then the four classes are classified and segregated with the same region of interest. Afterward, the same class-based areas are segmented, and these four class-based regions are next set as the next input of the next CNN layer with the clusters.

2. Literature review

James and Nordby (2002) defined forensic science as the function of the science discipline to related and help according to government laws. An elementary law of forensic science is an illegal act. Further, it is a human being initiated happenings and produces a record for that. The records are conversely flawed; also, it may be their actions, and the actions may locate in the movement that will produce interactive results that may change the atmosphere. The evidence may help them to get the results like relocated or damaged down, marks are made, and resources are transformed (Caddy, 2001). Forensic analysis is typically executed in the laboratories through exercises, but it consumes time, and human energy is very large. These days the number of datasets is available for the execution; it increases the practices easier, and it develops more computational intelligence techniques (Mumford, 2009). These computational intelligence techniques are extraordinarily significant in the process of accurate automatic prediction and fast analysis. Popescu and Farid (2004) proposed a technique regarding the geometric tool for the uses of the distorted photograph in a digital forensics perspective. While in digital forensics was introduces approximately for quite a few decades, but nowadays itself it is a juvenile science. Moreover, for doing any research or work, the review articles and the intellectual kinds of literature are very important. On behalf of every scientific research, there are many articles that are developed, but the contents creativeness is very small; however, scientific research is growing (Kessler, 2012). Based on the forensic nature or area, the evidence might be changed. In forensic, there is much evidence available where few of them are fibers, paint, glass, soil, fingerprints, and footprints (Houck, 2003).

Data mining and soft computing have many applications in the field of digital forensics. In this work, take account of identify the correlation in forensic data organization, discover and categorization of forensic data keen on to some group base on their match classification. The match group formations are done through the group location of clusters and then discover a pattern in a dataset with the purpose of prediction or forecasting (Abraham, 2006). Although this method is idyllic in data grouping, classification, clustering, and forecasting at the same time, it is the handiest technique for data visualization (Fayyad et al., 2003). Data visualization enables the digital investigator to locate vital information in the area of interest quickly and powerfully. Additionally, data visualization be capable of guide the digital investigator in the direction of digital evidence recovery with a more resourceful and successful style (Pernkopf, 2004).

In this paper, the forensic analysis carried out wild animal findings with their manifestation similarity match based. Our work aims to make available an experimental based follow a line of investigation on wild animal image segmentation and boundary detection. In this work, we used a collective dataset of 12,000 hand-labeled segmentation images of 1,000 Corel images from 30 human beings subject. The segmentation step executed with two steps as color and grayscale image processing. The dataset contains the half-color image; the other half is a grayscale image. This dataset contains a total of 500 images. In these images, some of the images are grayscale images, and the others are color images. For this work, the experimental setup of the whole images is separated into a training set as well as a test set (Agrawal et al., 1993).

The validation setup the whole images are separated into a training set for 300 images as well

as a test set for 200 images. Then the generated labels are compared with the ground truth labeling for a compartment of these images. It is beneficial for the researchers in the way of comparisons and gives some new ideas for budding new boundary detection algorithms (Bhat et al., 2010). The best way of proving our knowledge through competence was conducted by many datasets benchmark for their fame and new idea developments. This benchmark has been displayed in their implementation portal for boundary detector algorithms and steps alongside code, which is designed for executing the benchmark dataset. They have some dedicated team members to maintain the benchmark result in the fortitude of supportive scientific development (Grajeda et al., 2017).

3. Dataset

The forensic analysis carried for wild animal findings with their manifestation similarity matches basis was done in this work. Wild animal image segmentation and their boundary detection is the main scope of this work. The main goal of our work is the image segmentation and boundary detection based investigations over the natural images. The dataset which is used here is a group dataset of hand-labeled segmentation images of 12,000 pictures and the Corel image of 1,000 pictures commencing thirty dissimilar human beings image. Then the segmentation pace is performed through the color as well as the grayscale image processing. In the segmentation, the used dataset consists of a half-color image and a remaining grayscale image. Our dataset contains a set of 500 images within those images, many images are grayscale images, and the remaining are the color images. The proposed experimental arrangement of the entire dataset images is alienated into the training set as well as a test set. Here, the validation setup is estranged into a training set with 300 images, as well as a test set for 200 images. Then the validation process generated labels are compared with the ground truth labeling images in support of the image comparisons. The ground truth labels are essential in the way of comparison result production process for the researchers. Also, these comparisons will provide some new thoughts for upcoming boundary detection algorithms.

4. Methods and materials

Digital forensics follows a line of investigations which includes many stages. During digital forensics work, the initial task is data collection. Following the data collections, the output values have been predicted from the input data. So many researches are carried out with the supervised learning-based techniques, but the optimum output has not been obtained. In order to get better performance accuracy, a data mining based digital forensics task is proposed. A forensics dataset images are classified with class-based 3 layer CNN classifiers, 5 layer CNN classifiers, and 7 layer CNN classifiers. The validation results are calculated with the segmentation area prediction and efficiency. A fivefold cross performance with the fifty and hundred iterations for CNN algorithms for getting a reliable result. Then the comparison results are made with the used three algorithms.

The data selected for the experiments are the 500 natural images with all their ground truth human interpretations and their benchmark values. The datasets are split keen on dislodging training set, validation data, and the test subsets. Sequentially to defend the reliability of the valuation and acquire an expressive and reasonable assessment of results compared with the existing methods are made. For the testing and validation process, the following strategies may follow.

Training is the most important process in this work; the training process of this work has been made with train value. Every bit of the learning process through the CNN process, then the boundaries fine-tuning and their models' selections are made utterly with the training and validation subsets of the input data. Following the training process, the testing is done; the algorithm must execute simply with the unchanging parameter on the test subset of the dataset. The input image and ground truth segmentation values of the test set are not being used to modifying the algorithms. It is just for check and plots the evolutionary results. After the testing and training process, all the evaluation results are plotted. Evaluation results are made based on the test subset with the benchmark value. Consecutively in the direction of measure of performance aspects of the contour detection and the segmentation algorithm prediction results in the BSDS500 provide a matching set of estimation methods. Just describe all the score and curve return by the valuation characters' boundary Bench or the contour detection methods or all Bench or the segmentation methods. To construct the proposed result to compare with other existing methods was evaluated with the original BSDS300. It is functioning with the repeated above three steps. Excluding, make sure to the next two steps, i.e., train simply by train subset of the dataset and test simply by validation subset.

This work aims to provide digital forensics research based on image segmentation and boundary detection. In order to promote or improve the performance of the data mining approach for digital forensics tasks with deep learning techniques, the following resources are provided. The experimental plan has been done with fivefold cross-validation with fifty repetitions for deep learning algorithms in order to obtain consistent results. Matching accuracy values for the next to next pixels in the classes are calculated with the class-based predicted labels. There are four classes assigned on CNN, and the four classes are segmented and separated with the same region of interest. Then the same class-based region of interests is segregated, and these four class-based regions are next given to

CNN with the clusters. Further, the comparison results are made with the used three algorithms.

5. System model

The proposed overall architecture for the digital forensics segmentation method is shown in Fig. 1. This paper concentrates on the classification and segmentation of a nature image. Fig. 1 shows the system architecture of the proposed system. The proposed method consists of four stages named noise removal, feature extraction, feature extraction based classification of the image, and segmentation. Image segmentation is an essential pre-processing tread in a complicated and composite image dealing with the algorithm in nature images. Nature image detection is done initially by some other tests, but all are a costly and time-consuming process. Hence, a new nature image classification system is required for early classification and categorization of animals. In this work, the proposed system consists of pre-processing, feature formulation, and the corresponding algorithm. Then the segmentation is carried out based on the class-based features.

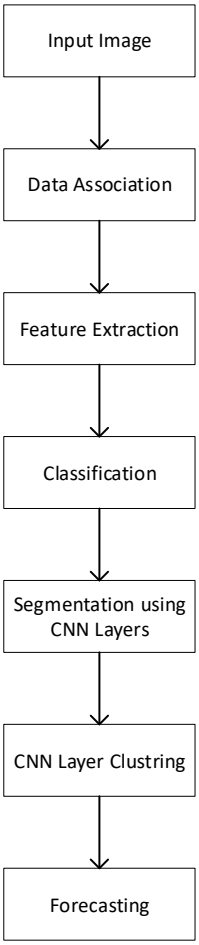

Fig. 1: Proposed overall architecture

Computer forensic tools in the current situation are not perfect, based on the following responsibilities. The first one is data association, which is used to identifying the correlation among the data. The second one is classification, which is used to discover and categorization data into a group based on data similarity. Then the clustering is to discover and visualize the present group of proofs unknown in the past or missing disregarded. The last one is forecasting, which is used for discovering patterns and data that may lead to reasonable predictions. The matching accuracy values for the next to next pixels in the classes are calculated with the class-based predicted labels. There are four classes assigned on CNN, and the four classes are segmented and separated with the same region of interest. Then the same class-based region of interests is segregated, and these four class-based regions are next given to CNN with the clusters.

6. Processing steps

In this section, pre-processing, feature extraction, histogram equalization, Gaussian filter, and CNN clustering are discussed.

6.1. Pre-processing

Nature images become corrupted by noise during image transmission and image digitization during the process of imaging. Pre-processing is necessary to remove such noises like unwanted pixels and also convert the heterogeneous image into a homogeneous image. Any filter will remove the noise in an image but also will corrupt minute details of the image. We adopt anisotropic diffusion filter for pre-processing of nature images as it removes noise and preserves the edges. This method helps to perform denoising and good for noises like shot or impulse noise.

6.2. Feature extraction

In this section, our plan to develop a class feature-based nature image segmentation method that improves nature image segmentation and classification. Common learning algorithm like boosting method helps to classify complex image texture and improve accuracy. CNN classifier yields a highly accurate component classifier for constructing a strong classifier as a linear combination of simple classifiers

6.3. Histogram equalization

Histogram equalization is a method in which the histogram of an image is obtained, and its contrast is adjusted. The equalization technique can be applied to the images that have backgrounds and foregrounds with different intensities. It is also known as the spatial domain method.

6.4. Gaussian filter

Gaussian filters are developed to avoid overshoot of step function input while reducing the rise and fall period. This temperament is especially much

correlated to the actuality to the Gaussian filter has the minimum possible group delay.

6.5. CNN clustering

Data clustering builds a supervised statistics model commencing of the data. Then the collected data instances group are collectively based on correspondence class-based. The processes of the clustering system showing Fig. 2 in a large dataset of multidimensional images are very complicated; all these are overcome by the CNN methods. After the segmentation process, the clustering is the best one to group data instance keen on clusters of significant interest. Then the performances are evaluated for the models and detect forensics. Here the class-based selected clusters is a key process to the analysis of data generation area in the digital forensic investigation.

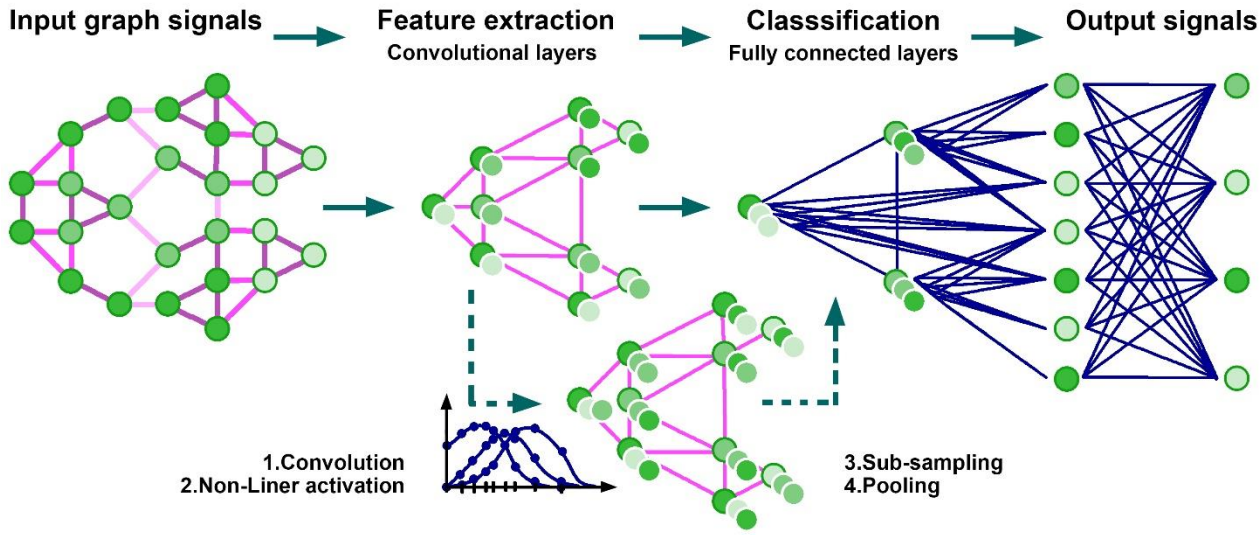

Fig. 2: CNN clustering process

Our concentration is scrutinizing individual data instance those are not grouped obviously within the cluster groups for forensic verification. In this work, three layer-based CNN algorithm for the collection of clusters. Three layer-based CNN algorithm for the collection algorithm takes n number of clusters to determine. The 'I' is the input constraint as well as subdivisions' is specified with a group of 'k' substance keen on 'n' cluster. Accordingly, the resultant intra-cluster comparison was large at the same time as the inter-cluster comparison are small. Euclidean distance calculation was essential in the direction of allocating instance towards the calculation of the clusters. Cluster comparisons are calculated with the mean rate of the core in a cluster.

7. Experimental results

The circulated training is performed with 50, and 20 iterations for the whole nature image segmentation mask of every worker nodule received one-quarter of their batch. To process, along with the parameter server, rationalized the gradient synchronously. The numbers of color, as well as greyscale images of pixel size 12×12, 24×24, 36×36, are used in this work for the validation process. In machine learning architectures, the computational capacity was restricted. The technique consequently was not scalable to large scale images. So, in order to overcome the computational capacity and reduce the time consumption in this works, the deep learning techniques are proposed. The model contained 3, 5, 7 layers' architectures.

In the proposed CNN layer in Fig. 3, the first layer is a convolutional layer of kernel size of 4×4, a stride of 1 and 6 kernels in overall. So the input image of size 36×36×1 gives an output of 24×24×4. The second layer is a pooling layer with 1×1 kernel size, a stride of 2×2, and 6 kernels in overall. In this network pooling layer is a different one. The input pixel values in the amenable are added up, and after that, the values are multiplied to a single value per filter parameter. Finally, these results were added to a single value per filter trainable bias. The output value was achieved by applying the sigmoid activation. Therefore, the input value from the previous layer of size 24×24×4 gets sub-sampled to 12×12×4. Total parameter values in this layer are equal to the sum of one trainable parameter and one trainable bias. Then this value is multiplied with the 4 filter kernel value.

Layer 3 is similar to Layer 1. It has a convolutional layer of kernel size of 4×4, a stride of 1 and 6 kernels in overall. So the input image of size 36×36×1 gives an output of 24×24×4. This layer is a convolutional layer with the same configuration apart from the 12 filters instead of 4. So, the input from the previous layer of size 12×12×4 gives an output of 8×8×12 the extracted feature map is 416. Layer 4 is the same as layer 2; this layer is pooling layers with 12 filters this time around the output are passed through sigmoid establishment function. The

input of size 8×8×12 from the previous layer gets sub-sampled to 4×4×14. The total parameters are 30.

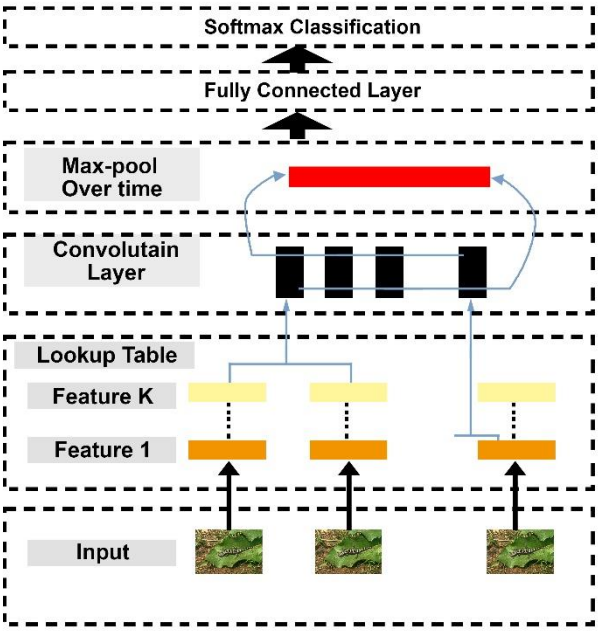

Fig. 3: Proposed CNN architecture

In layer 5, we got a convolutional layer with 4×4 kernel size and 100 filters. There is no need even to consider strides as the input size is 4×4×14 so we will get an output of 1×1×100. Total parameters in layer=4×4×100=1600. The sixth layer had a dense layer of 74 parameters. So, the input of 100 units is converted to 74 units. Total feature maps are calculated by 74×100+74=74074. A unique activation function is used here. In the output layer of 8 units of dense layers were used. Total predicted feature maps of obtained size are as 74×8+8=600. The problem while using a convolutional layer network was the used loss function, so in order to avoid it by using cross-entropy loss function with softmax activation in the last convolutional layer.

The distributed training is performed with 100, and 50 iterations for the segmentation mask of every worker nodule received. One-quarter of the batch to process along with the parameter server rationalized the gradient synchronously. The proposed process with the original CNN model is halved the number of feature maps, i.e., 12→24→36→48→112 used. Feature maps of half model create the same dice scores of the full model though getting a better time to train the similar padding lying on all convolutional layers. It allowed getting free of cropping function throughout the concatenation of feature maps. Fig. 4 CNN's building blocks are convolutional layers; these convolutional layers are load up on top of each other to make a step of features. Each convolutional layer is set to the following five filter kernels sizes like 2×2, 2×2, 2×2, 2×2, and 2×2.

For first convolutional layer 64 feature maps of size 237×237 as C1, acquires to related image modalities patches of MRI as inputs. After 5 convolutional layers, the pooling function takes consign to entail the highest feature value over sub-windows at an interval of every feature map; it introduces the property of invariance to local translation to retreat the size of the corresponding feature map. Individually the weights of the predicted three scales are learning, and then the predicted top three scales joint together to produce a new three-pathway network structure. Proposed Multiscale-Multimodality CNN's owe to these dissimilar tasks specific hierarchies of features extracting procedures in addition to every output of three pathways are combined in full connection layer as the input of final classification of the input image patch of the center pixel. In patch classification, three pathway features are arranged in one dimension and utilize cooperatively for classification in the next three-layer module by the vector illustration of tumor image.

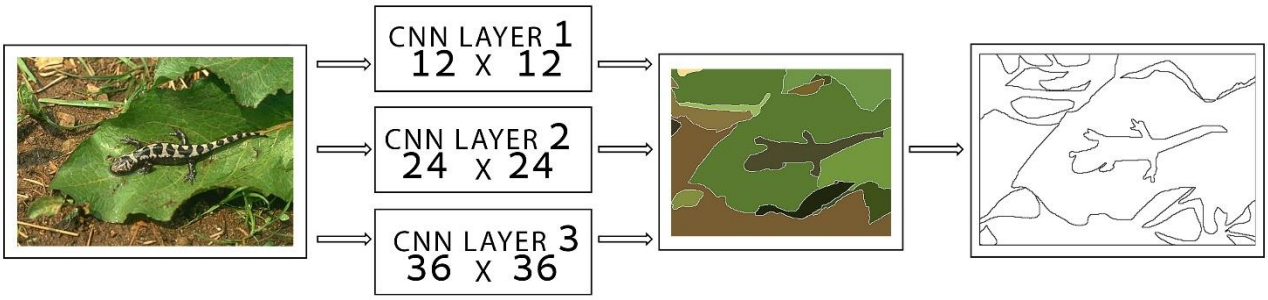

Fig. 4: CNN block diagram

7.1. Validation technique

The experimental plan has been done with fivefold cross-validation with fifty repetitions for deep learning algorithms in order to obtain consistent results. Matching accuracy values for the next to next pixels in the classes are calculated with the class-based predicted labels. The most important idea is that the fivefold cross-validation method was dividing the complete dataset into five portions of similar sizes. Full data set in five partitions are worn at the same time as a test set, and the leftover is worn as a train set. Stratification subject to with the purpose of the partition maintains the class allocation of the samples in the region equal in the original data set. The proposed algorithm has executed twenty times, and since we have fivefold.

7.2. Data mining server

Data mining server is necessary in the direction of a data mining classification, and an ideal world contains a set of well-designed modules. These well-designed modules used for the data description, organization, clustering analysis, classification, and development and divergence investigations. In this work, there are three steps that are used to analyze the dataset. The first step is to classify the image classes by the CNN supervised classification algorithm. Then the data are separated based on class clusters the image classes then the grouping is done with CNN layers algorithm. Finally, the classification algorithms are used to verifying the visualization pattern of the data instances.

7.3. Clustering

A supervised statistics representation origination of the data is generated by the data clustering. After that, the composed data instance sets are cooperatively collected base on correspondence class-basis. The clustering classification with the large dataset of multidimensional images are very difficult to process, but it can be done by the deep learning process. Here the clustering classification of multidimensional images is conquered through the CNN methods. The clustering process is the best one to group the data instances subsequent to the segmentation process enthusiastic on clusters of significant attention. Finally, the performances are evaluated on behalf of this model and detect the needful forensics. By the side of this work, the class-based cluster selection is the best solution to the investigation of data production region in the digital forensic explorations. Our main aim is to scrutinize the individual data instance which is not grouped perceptibly inside the cluster groups, which is designed for forensic verification. In the Euclidean distance $S(i)$ of clustering calculation in CNN c (i) is the average distance between the cluster and x (i) is the distance between the next to next clusters.

$$S(i) = \frac{c(i)-x(i)}{\max(c(i),x(i))} \quad (1)$$

In this proposal, a layer-based CNN algorithm is used for the collection of clusters. There are three different layer-based CNN method intended in the cluster collection. In this algorithm, it takes n number of clusters to determine the collection of clusters. The formula for standard deviation is:

$$SD = \sqrt{\frac{\sum|x-x\prime|}{n}} \quad (2)$$

Here 'I' is the input control key as well as groups' which is specific with a group of 'k' gist keen on 'n' cluster. Therefore, the resulting intra-cluster evaluation was huge at the same time the inter-cluster comparison is little. Euclidean distance computation is necessary for the direction of allocating illustration towards the computation of the clusters. Finally, the cluster comparison is designed through the mean rate of the central part in a cluster.

From Table 1, predicted results, the increasing layer will provide the best layer for classification and clustering. The matching accuracy values for the next to next pixels in the classes are calculated with the class-based predicted labels. There are four classes assigned on CNN, and the four classes are segmented and separated with the same region of interest. Then the same class-based region of interests is segregated, and these four class-based regions are next given to CNN with the clusters. Then the clusters are the comparison key for digital forensics. In the comparison part, the two dissimilar inputs are classified, and the output is zero in deep learning.

Table 1: CNN layer comparison

Methods	Matching Accuracy	Cluster Mean Distance	Standard Deviation	Efficiency
3 Layers CNN	85.16	4.56	6.32	73%
5 Layers CNN	86.21	5.46	6.58	76%
7 Layers CNN	90.46	5.92	6.92	81%

7.4. Classification

The input data are fed as input to the classifier. We have selected three-layer CNN methods. In the proposed CNN layer, the first layer is a convolutional layer of kernel size of 4×4, a stride of 1 and 6 kernels in overall. So, Fig. 5, the input image of size 36×36×1 gives an output of 24×24×4. The second layer is a pooling layer with 1×1 kernel size, the stride of 2×2, and 6 kernels in overall. In this network pooling layer is a different one. The input pixel values in the amenable are added up, and after that, the values are multiplied to a single value per filter parameter. Finally, these results were added to a single value per filter trainable bias. The output value was achieved by applying the sigmoid activation.

Fig. 5: Segmentation results

Therefore, the input value from the previous layer of size 24×24×4 gets sub-sampled to 12×12×4. Total parameter values in this layer is equal to the sum of one trainable parameter and one trainable bias; then, this value is multiplied with the 4 filter kernel value. The standard way of predicting the error rate of a classifier given a single, fixed stratified data is to use fivefold cross-validation. The cross-validation technique is adopted in cases when the amount of data for training and testing is limited.

7.5. CNN layer clustering

Data clustering builds a supervised statistics model commencing of the data. Then the collected data instances group are collectively based on correspondence class-based. The processes of the clustering system in a large dataset of multidimensional images are very complicated; all these are overcome by the CNN methods. After the segmentation process, the clustering is the best one to group data instance keen on clusters of significant interest. Then the performances are evaluated for the models and detect forensics. Here, the class-based selected clusters is a key process to the analysis of the data generation area in the digital forensic investigation. Fig. 6 shows the linear cluster efficiency between the test and training set.

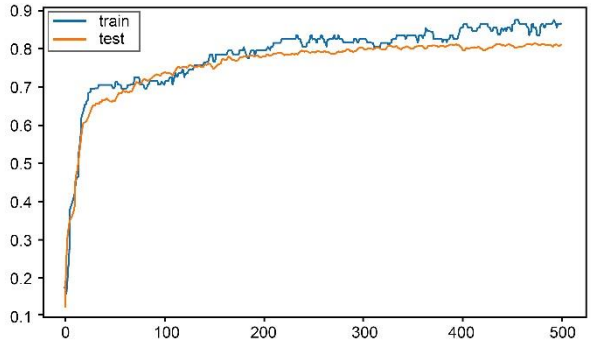

Fig. 6: Linear cluster efficiency between test and training set

Fig. 6 shows the linear cluster efficiency between the test set and the training set. From that, it is clear that the training set within the cluster provides the best output performance results. This layer cluster results you can see in Fig. 7. Our concentration is scrutinizing individual data instances that are not grouped into the cluster groups for forensic verification. In this work, three layer-based CNN algorithm for the collection of clusters. Three layer-based CNN algorithm for the collection algorithm takes n number of clusters to be determined. The 'I' is the input constraint as well as subdivisions' is specified with a group of 'k' substance keen on 'n' cluster. Accordingly, the resultant intra-cluster comparison was large at the same time as the inter-cluster comparison are small. Euclidean distance calculation was essential in the direction of allocating instance towards the calculation of the clusters. Cluster comparisons are calculated with a mean rate of the core in a cluster.

The matching accuracy values for the next to next pixels in the classes are calculated with the class-based predicted labels. There are four classes assigned on CNN, and the four classes are segmented and separated with the same region of interest. Then the same class-based region of interests is segregated, and these four class-based regions are next given to CNN with the clusters.

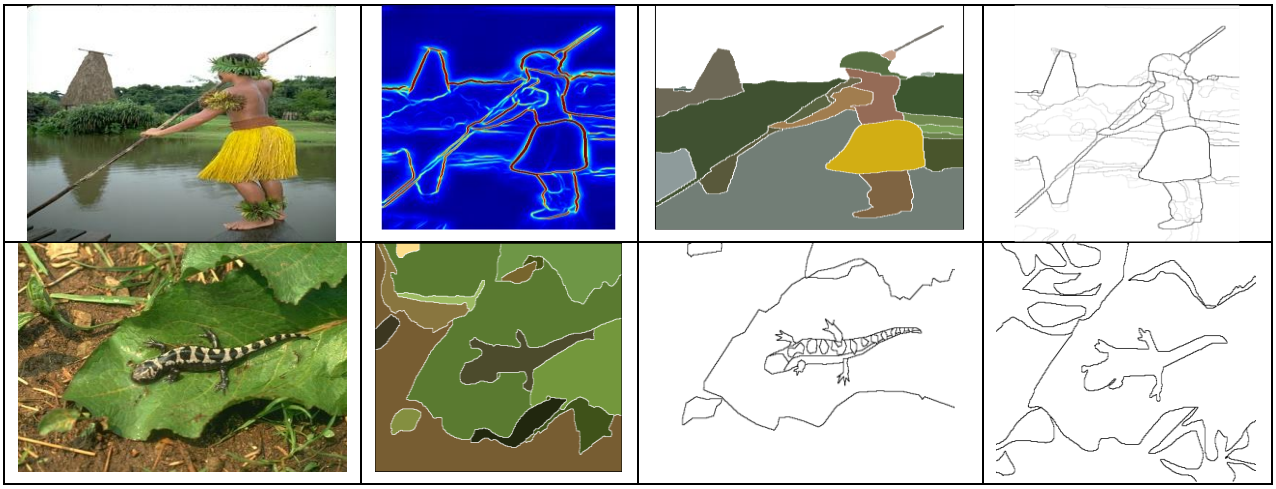

Fig: 7: Layer cluster results

7.6. Digital forensics

Digital forensics research includes many stages. In digital forensics, the first task is collection; after that, the output has been predicted with that hidden data. Some authors proposed several supervised machine learning techniques which have not obtained much better results. So, in order to improve the performance of the work, a data mining approach for digital forensics task with deep learning techniques are proposed. This paper performs an investigational work on a forensics dataset task for class-based classification, including several types of methods such as 3 layer CNN classifiers, 5 layer CNN classifiers, and 7layer CNN classifiers. The classifiers have been evaluated with classification performance and accuracy. The followed experimental design has been fivefold cross-validation with fifty repetitions for deep learning algorithms in order to obtain consistent results. The matching accuracy values for the next to next pixels in the classes are calculated with the class-based predicted labels. There are four classes assigned in CNN, and the four classes are segmented and separated with the same region of interest. Then the same class-based region of interests is

segregated, and these four class-based regions are next given to CNN with the clusters.

8. Conclusion

In the past, digital forensic, with its exploration techniques, are a lane to the data recovery as well as the examination of different investigation techniques. It is a line of investigation which includes many stages. In this, the foremost assignment is data collection later than that the outcome amount produced predicted with the dataset. The supervised machine learning technique is proposed earlier but has not been obtained optimum results. So, with the purpose of performance evaluation, a data mining approach for digital forensics tasks with deep learning techniques is proposed. This work performs a natural image forensics dataset task based on the class-based classification together with many methods such as 3 layer CNN classifiers, 5 layer CNN classifiers, and 7layer CNN classifiers. The classifiers' accuracy has been evaluated with the categorization presentation. Fivefold cross-validation with fifty repetitions for deep learning algorithms in order to attain reliable outcomes. A statistical analysis has been conducted in order to compare each pair of algorithms.

The data that were selected for the experiments are the 500 natural images with all their ground truth human interpretations and their benchmark values. The datasets are split keen on dislodging training set, validation data, and the test subsets. Training is the most important process in this work; the training process of this work has been made with train value. Every bit of the learning process through the CNN process, then the boundaries fine-tuning and their models' selections are made utterly with the training and validation subsets of the input data. Following the training process, the testing is done; the algorithm must execute simply with the unchanging parameter on the test subset of the dataset.

The input image and ground truth segmentation values of the test set are not being used to modify the algorithms. It is just for check and plots the evolutionary results. After the testing and training process, all the evaluation results are plotted. Evaluation results are made based on the test subset with the benchmark value. Consecutively in the direction of measure of performance aspects of the contour detection and the segmentation algorithm prediction results from the BSDS500, provide a matching set of estimation methods. Just describe all the score and curve return by the valuation characters' boundary Bench or the contour detection methods or all bench or the segmentation methods. The matching accuracy values for the next to next pixels in the classes are calculated with the class-based predicted labels. There are four classes assigned on CNN, and the four classes are segmented and separated with the same region of interest. Then the same class-based region of interests is

segregated, and these four class-based regions are next given to CNN with the clusters.

Acknowledgment

The author thanks King Abdulaziz University for this work.

Compliance with ethical standards

Conflict of interest

The authors declare that they have no conflict of interest.

References

Abraham T (2006). Event sequence mining to develop profiles for computer forensic investigation purposes. In the 2006 Australasian Workshops on Grid Computing and E-Research, Australian Computer Society Inc., Hobart, Australia, 54: 145-153.

Agrawal R, Imieliński T, and Swami A (1993). Mining association rules between sets of items in large databases. ACM SIGMOD Record, 22(2): 207-216. https://doi.org/10.1145/170036.170072

Beebe NL and Clark JG (2005). A hierarchical, objectives-based framework for the digital investigations process. Digital Investigation, 2(2): 147-167. https://doi.org/10.1016/j.diin.2005.04.002

Bhat VH, Rao PG, Abhilash RV, Shenoy PD, Venugopal KR, and Patnaik LM (2010). A data mining approach for data generation and analysis for digital forensic application. International Journal of Engineering and Technology, 2(3): 313-319. https://doi.org/10.7763/IJET.2010.V2.140

Caddy B (2001). Forensic examination of glass and paint: Analysis and interpretation. CRC press, Boca Raton, USA. https://doi.org/10.1201/9780203483589

Carrier B (2003). Defining digital forensic examination and analysis tools using abstraction layers. International Journal of Digital Evidence, 1(4): 1-12.

Casey E (2009). Handbook of digital forensics and investigation. Academic Press, Cambridge, USA. https://doi.org/10.1016/B978-0-12-374267-4.00004-5 **PMid:20881312**

Conlan K, Baggili I, and Breitinger F (2016). Anti-forensics: Furthering digital forensic science through a new extended, granular taxonomy. Digital Investigation, 18: S66-S75. https://doi.org/10.1016/j.diin.2016.04.006

Fayyad UM, Piatetsky-Shapiro G, and Uthurusamy R (2003). Summary from the KDD-03 panel: Data mining: The next 10 years. ACM SIGKDD Explorations Newsletter, 5(2): 191-196. https://doi.org/10.1145/980972.981004

Grajeda C, Breitinger F, and Baggili I (2017). Availability of datasets for digital forensics–And what is missing. Digital Investigation, 22: S94-S105. https://doi.org/10.1016/j.diin.2017.06.004

Houck MM (2003). Trace evidence analysis: More cases in forensic microscopy and mute witnesses. Elsevier, Amsterdam, Netherlands.

James SH and Nordby JJ (2002). Forensic science: An introduction to scientific and investigative techniques. CRC Press, Boca Raton, USA.

Kessler GC (2012). Advancing the science of digital forensics. Computer, 45(12): 25-27. https://doi.org/10.1109/MC.2012.399

Mumford CL (2009). Synergy in computational intelligence. In: Mumford CL and Jain LC (Eds.), Computational intelligence: 3-21. Springer, Berlin, Germany. https://doi.org/10.1007/978-3-642-01799-5

Pernkopf F (2004). Detection of surface defects on raw steel blocks using Bayesian network classifiers. Pattern Analysis and Applications, 7(3): 333-342. https://doi.org/10.1007/s10044-004-0232-3

Popescu AC and Farid H (2004). Statistical tools for digital forensics. In: Fridrich J (Ed.), International workshop on information hiding, Berkeley, USA: 128-147. Springer, Berlin, Germany. https://doi.org/10.1007/978-3-540-30114-1_10

Reith M, Carr C, and Gunsch G (2002). An examination of digital forensic models. International Journal of Digital Evidence, 1(3): 1-12.

Shahraki AS, Sayyadi H, AMRI MH, and Nikmaram M (2013). Survey: Video forensic tools. Journal of Theoretical and Applied Information Technology, 47(1): 98-107.