Optimized deep learning model for spatio-temporal detection and localization of object removal video forgery with multiple feature extraction

Lakshmi Kumari Ch (lakshmikumari96@yahoo.com)
Koneru Lakshmaiah Education Foundation
PRASAD K.V
Koneru Lakshmaiah Education Foundation

Research Article

Keywords: Video forgery detection and localization, GMM model, Pre-processing, multiple features, ResNet152V2, Bi-GRU, Improved remora optimization

Posted Date: May 24th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1641193/v1

Optimized deep learning model for spatio-temporal detection and localization of object removal video forgery with multiple feature extraction

Lakshmi Kumari Ch [1*], K.V.PRASAD [2]

[1] Research Scholar, Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur-Dt, Andhra Pradesh.

[2] Associate Professor, Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur-Dt, Andhra Pradesh.

[1*] lakshmikumari96@yahoo.com, [2]prasad_kz@yahoo.co.in

Abstract

Video forgery (VF) is an approach for manipulating the fake videos by modifying, coordinating or generating new contents among the video sequence. The identification of this type of video forgery is complex. For extracting multiple features, developing a novel approach still remains major challenge in this area. In this work, an optimized deep learning (DL) model for the video forgery detection (VFD) and localization with multiple feature extraction is proposed. Initially, key-frame extraction takes place with the aid of Gaussian mixture model (GMM) to extract frames from the forged videos. Then, pre-processing stage is manipulated for the conversion of RGB frame into grayscale image. To study the nature of the forged videos, there is a need of extracting multi-features from the pre-processed frames. In our proposed study, SURF, PCA-HOG, MBFDF, COA and PRG features are extracted. The dataset used for proposed work is collected from REWIND of about 80 forged and authenticated videos. With the help of DL approach, video forgery can be detected and localized. Thus, this research mainly focus on the detection and localization of forged video based on ResNet152V2 model hybrid with Bi-GRU to attain the maximum accuracy and efficiency. The performance of this model is finally compared with existing approachesin terms of accuracy, precision, F-measure, sensitivity, specificity, FNR, FDR, FPR, MCC and NPV. The proposed methodology assures the performance of96.17% accuracy, 96% precision, 96.14% F-measure, 96.58% sensitivity, 96.5% specificity, 0.034 FNR, 0.04 FDR, 0.034 FPR, 0.92 MCC, 96% NPV respectively. Along with is, the mean square error (MSE) and peak-to-signal-noise ratio (PSNR) for GMM model attained about 104 and 27.95 respectively.

Keywords: Video forgery detection and localization, GMM model, Pre-processing, multiple features, ResNet152V2, Bi-GRU, Improved remora optimization.

1. Introduction

With a fast advancement of science and technology, video processing and computing play a dominant role in day today's life.Nowadays, the digital video contents are easily alter by anybody with the aid of powerful tools like adobe premiere, adobe after effects and apple final cut pro etc. [1]. 'Video forgery' is a method of developing modified or fake videos by combining altering or creating new video [2].The videos based on alternate objects that spreads all over the country through social media that leads to serious black mark in social security among the people [3]. In recent world, many researchers deeply investigate the elimination of video forgery detection. But still the authenticity of these digital videos is remarkable and needs to be verified with advancements in growing technology [4].

The term ‘object forgery’ in videos is a general video altering process that can add new objects along the video sequence or it can eliminate the original ones [5]. For contrasting the image copy-move forgery (CMF) approaches, object based VFD is truthfully a difficult work. The forgery videos are sub divided into three different types: (i) object-based tampering, (ii) frame based tampering, (iii) spatiotemporal tampering [6]. On the contrary, if image CMF algorithms are implemented for video forgery detection that results in high computational cost. However, the techniques for image forgery detection cannot be directly applied for the video forgery detection [7]. For overcoming this issue, the secular correlation among the video frames must be considered as important factor for reducing the complexity of detection in video forgery [8].

To overcome this challenging task, multiple video forgery algorithms have been implemented in past years. Some of the existing approaches proposed to detect the pixel-similarity among various video frames. For the first, multimedia signal processing (MMSP) based algorithms is introduced to detect the image forgery defend and video forgery defend with the aid of correlation analysis for video frames and video blocks [9]. But, this method is highly suffered due to high noisy in the final output and there is lack of filtering techniques. To overcome this issue, another passive implementation takes place for the detection and localization of video forgery [10]. This method utilizes some extraction techniques to fetch the specific features based on spatiotemporal coherence analysis. Here, the classifications are employed using machine learning algorithms. But, there is limitation is that highly time consuming process with high error in the outcome. To solve this multiple issues, passive CMF detection in videos using SIFT method [11]. This method uses SIFT approach for withdrawing the features from the frames. Finally, for finding the spatial CMF K-NN matching technique is used. However, all the aforementioned techniques mostly hang on traditionally designed features generated from forgery and pristine video sequences.

Nowadays, with the rapid advancements in technology, deep learning model play a dominant role in several applications such as frame forgery detection, object forgery detection and localization etc. in the field of computer vision [12]. DL approach have the capability of withdrawing features with high dimensions and generate fascinating outcome [13]. In today’s life, deep learning approach is used in enormous field like image CMF, image handling detection, camera model identification, steganalysis, video modification forgery and so on [14].

Many researchers and scientists are interested with the deep learning based CNN model but the task done by this model is still a challenging work. The CNN model has the multiple limitation as it needs larger amount of training data and it makes the process more complex and increases the time consumption [15].For the identification of forgery videos, a newly advanced methodology is highly required for attaining better accuracy. Although peripheral researches were conducted in this work, there still arises major limitations in identifying the forgery videos. These kind of major drawbacks motivate us to propose optimized deep learning model for detecting and localizing the forgery videos. So, a DL based residual network (ResNet) model is introduced to solve the aforementioned issues. The multi-objective strategies are introduced and it is perfectly useful for many applications in the research field. Hence, this multi-objective strategies are highly apt with the network model to efficiently detect and localize the video forgery with the efficient accuracy instantly.

In this paper, we propose a newly advanced methodology for object based video forgery detection and localization based on DL technique. In this work, we have proposed the remora algorithm based ResNet152V2 with bi-directional gated recurrent unit (Bi-GRU) model for

detecting the object removal forgery and to localize the forged region.Due to its fast training operation, significant learning rates, our proposed model shows highly efficient outcome in extracting the multi features from the video get forged. The ResNet152V2 is a pre-trained model to accelerate the training process and produces high accuracy rapidly. With the aid of this method, the deep learning models can develop efficient and flexible models that shows higher accuracy. The Bi-GRU present here, is one of the RNN architecture that have the ability to predict the information irrelevant to truthfulness for the longer time without destroying it. The Bi-GRU has multiple advantages like very simple implementation, easy modification and takes less time to train the model, which is more efficient and suitable for many application. In addition to this network model, the Gaussian mixture model (GMM) is used for extracting the frame from the forged videos. The extracted frames are then subjected to pre-processing stage and this process helps in converting RGB frames into gray scale images. Finally, the frame are added to multiple features for extracting important features from the video frames. Our proposed work undergoes the following major contributions and it is discussed below:

- This innovative research work mainly aims to optimize the DL model for the detection and localization of object removal VF with multi-feature extraction undergoes four stages includes key frame extraction, preprocessing, feature extraction and detection.
- For the efficient operation, the video is initially converted into frames. To perform this, GMM model is introduced and it is efficient in extracting foreground and background images from the entire frame.
- To eliminate the unwanted ripple and to increase the quality of the frames, pre-processing stage is emphasized. In this stage, the videos with RGB frames are transformed into grayscale frames.
- To improve efficiency of the accuracy of the optimization algorithm (OA) and time reduction, feature extraction stage is introduced in this proposed work.
- To detect and localize the object removal forgery in videos, ResNet152V2 model is enhanced. This model helps to speed up the training process and avoid gradient problems in video frames.
- For fine tuning the hyper parameters and loss function reduction, improved remora optimization algorithm is emphasized in this work.
- The evaluation of the proposed model are with SULFA dataset and the performance of accuracy, precision, recall, F-measure, specificity, FPR, FDR, NPV, MCC, FNR and sensitivity are analyzed and compared with existing techniques.

The oncoming division are explained in detail as follows: division 2 describes the review of recently issued papers related to VFD. Division 3 manipulates the proposed methodologies. Division 4 describes the results and discussion. Division 5 demonstrates the conclusion of the proposed approach.

2. Related work:

Some of the recently published paper are surveyed below:

Yao et al. [16] had defined the DL for detection of object-based forgery in latest video. In this method, CNN was emphasized to extract maximum level dimensionality features from the source frames. Here, max pooling layer was employed to eliminate temporal redundancy between the video frames. For enhancing the residual signals from video forgery, high pass

filter layer was introduced. Before performing the training process, asymmetric data augmentation strategy was employed to find the differences among negative and positive image patches. The dataset used here was SYSU-OBJFORG of about 100 pristine video sequences and 100 forged videos. In experimental scenario, the maximum accuracy attained was about 89%. However, this method shows low resolution video sequence in the outcome.

D'Avino et al. [17] investigated the auto-encoder with RNN for VFD. In this method, auto-encoder studies the intrinsic model of the pristine frames during training phase. Here, RNN along with LSTM was emphasized to destroy the temporal dependencies of the video frames. Here, the feature extraction process undergoes three stages: using high pass filter to compute the residuals, residual quantization and co-occurrences of histogram computation. The experiment was carried out through GRIP and Hollywood camera work datasets. The tensor flow based adam learning algorithm was used for implementing this proposed work. In experimental scenario, the maximum accuracy attained was about 92%. However, this method was suffered due to high noise in the video frames due to lack of pre-processing stage.

Sasikumar et al. [18] studied the VFD using DL techniques and clustering algorithms. In this method, scalar invariant feature transform (SIFT) and mean shift clustering algorithms (MSCL) was enhanced to assemble the equal image frames from the frame extraction videos. Here, edge forecasting method was employed in each frames to avoid mistakes during video assembling. Here, SIFT algorithm was employed to detect different features from each image. For gathering the edge dependent on the video frames, segmentation method was employed. For clustering process, MSC algorithm was performed in each video frames. For experimentation, the dataset collected from ordinary CCTV video footages of about 100 forged videos. In experimental scenario, the maximum accuracy attained was about 93%. However, this method was suffered due to complex algorithm and time consuming process.

Kohli et al. [19] proposed the CNN aided localization of forged region in object-based forgery for HD videos. In this method, initially forgery frames were detected using temporal CNN and then the forged region was localized with the aid of spatial CNN. Here, the whole network process is run by using motion residual. For experimenting, SYSU-OBJFORG dataset was emphasized. Here, the binary classification such as doubled compressed frames and forged frames. The algorithm used for implementation was exponent Fourier moments (EFMs) for the detection of forged regions. The extracted EFMs was then optimized from the frame replicated region. The overall performance for this method was calculated on the basis of precision, recall, F-measure and accuracy. In experimental scenario, the overall accuracy attained was about 91%. However, this method was suffered due to high temporal noise in the video sequence.

Fadl et al. [20] introduced the CNN spatio-temporal features and fusion for surveillance VFD. In this method, inter frame forgeries like frame removal, frame addition and frame replication can be overcome using 2D-CNN based deep learning approach was established. For classification process, Gaussian RBF multi-class support vector machine (RBF-MSVM) was emphasized. For easy extraction of features from the frames spatiotemporal average (STP) fusion technique was given to each video frames. For experimentation, the datasets collected from SULFA, LASIESTA and VIRAT of about 200 forged and original videos. For feature extraction, VGG network was employed and help easy calibration of SSIM scores (scores of STP images). The overall performance of this method was calculated on the basis of sensitivity, specificity and accuracy. In experimental scenario, the maximum accuracy

attained was about 94%. However, this method was suffered due to high computational cost and time complexity.

Patel et al. [21] defined the optimized convolutional neural network based inter-frame forgery detection model – A multi-feature extraction framework. In this method, the original frames was pre-processed to demonstrate the quality of the image. Some of the features such as SURF, PCA-HOG, MBFDF, CAF, PRG and OFG were extracted. The extracted features were provided to optimized CNN with the aid of fine-tuned weights using hybrid approach. For efficient working, mayfly optimization was mathematically fused with black window optimization (MO-BWO). The entire experiment was analyzed using MATLAB software and dataset collected from REWIND and GRIP of about 80 forged and authenticated videos. The overall performance of this method was calculated using accuracy, precision, sensitivity, specificity, FNR, FDR, FNR and FPR. In experimental scenario, the maximum accuracy attained was about 85%. However, this method was shows poor accuracy due to high error in the outcome.

Soeleman et al. [22] proposed the adaptive threshold for moving object detection using GMM. In this method, Otsu algorithm and gray threshold act as the ground approach for performing the output based on MSE and PSNR. The Gaussian mixture model used here to extract the forefront and back front sequence from the source frame. The grey image was split into different regions automatically with the aid of otsu method. For reducing the noise in the foreground frame, morphological filters were evaluated. For experimentation, the human video dataset was utilized and performance was calculated on the basis of MSE and PSNR. In experimental scenario, the maximum PSNR and MSE attained was about 24.71 and 257.18. However, this method was suffered to low quality outcome due to high Gaussian noise.

Table 1 describes the comparison of existing techniques.

Author & reference	**Method**	**Performance**	**Merits**	**Demerits**
Yao et al. []	The deep learning for detection of object-based forgery in advanced video	Accuracy attained was about 89%.	Noise is highly reduced.	Low resolution video sequence in the outcome.
D'Avino et al. [17]	Auto-encoder with RNN for video forgery detection	Accuracy attained was about 92%	Easy implementation, fast output.	High noise in the video frames due to lack of pre-processing stage.
Sasikumar et al. [18]	VFD using DL techniques and clustering algorithms	Accuracy attained was about 93%	Low noise and high quality outcome.	Complex algorithm and time consuming process.

Kohli et al. [19]	CNN based localization of forged region in object-based forgery for HD videos	Accuracy attained was about 91%.	High outcome, easy implementation and fast execution.	High temporal noise in the video sequence.
Fadl et al. [20]	CNN spatiotemporal features and fusion for surveillance VFD	Accuracy attained was about 90%.	Easy implementation and low Gaussian noise	High computational cost and time complexity.
Patel et al. [21]	Optimized CNN based inter-frame forgery detection model – A multi-feature extraction framework	Accuracy attained was about 85%.	Low computational cost and fast execution.	Poor accuracy due to high error in the outcome.
Soeleman et al. [22]	Adaptive threshold for moving object detection using GMM	PSNR and MSE attained was about 24.71 and 257.18	Gaussian mixture model shows better efficiency compared to other methods.	Low quality outcome due to high Gaussian noise.

Table 1: Comparison of existing techniques.

Problem formulation

By analyzing the deep search in surveyed papers, the existing methods are highly suffered due to major drawbacks and these method shows less accuracy. In [16], the DL based detection of object-based forgery in advanced video was proposed. But the aforementioned model shows low resolution video sequence in the outcome. In [17], the author performed auto-encoder with RNN for video forgery detection. But this method shows high noise in the video frames due to lack of pre-processing stage. In [18], the author introduced video forgery detection using DL approach and clustering algorithm. But the MSCL algorithm was a complex process and takes huge time to process. In [19], CNN based localization of forged region in object-based forgery for HD videos was emphasized. However, the aforementioned method was highly prone to temporal noise in the video sequence. In [20], the author presented CNN spatiotemporal features and fusion surveillance VFD. But this method consumes large amount of time and also high computational cost. In [21], optimized CNN based inter-frame detection model with a multi-feature extraction framework was enhanced. However, the aforementioned method shows poor accuracy due to high error in the outcome.

In [22], the author proposed the adaptive threshold for moving object detection using GMM. But this method was highly due to low quality in the outcome because of Gaussian noise.

Only peripheral researches are emphasized for the detection and localization of object based video forgery using deep learning, machine learning and optimization techniques. The aforementioned survey paper are highly efficient with better classifying outcome, despite these techniques are high cost for performing the computation that is impractical in real-time applications. To deal with the efficient accuracy and fast performance, an efficient novel approach is proposed in our work. To the best of our awareness, our proposed approach is the primary research carried out in this field that can optimize the error and shows better accuracy efficiently.

3. Proposed video forgery approach

3.1 Proposed model

The video forgery detection aids to observe the verification of hardening by emphasizing the authenticity of the digital video evidence. In this method, deep learning based VFD and localization is proposed to detect the presence of authentication or forgery among video sequences. This research undergoes four major stages. The stages include (i) key frame extraction (ii) pre-processing (iii) extraction of multiple features (iv) detection of forged video. For processing the work, initially the frames must be collected from the video sequence. To perform this, Gaussian mixture model (GMM) play a dominant role in extracting key frames from the video sequence. After extracting the key frames from the video, pre-processing stage is enhanced. In this stage, the video converted RGB frames are transformed into grayscale frames.

The grayscale frames are given to extract multi-features for the detection and localization of VF. Multi-features such as speeded up robust features (SURF), principal component analysis, histogram of oriented gradient (PCA-HOG) feature, model based fast digit feature (MBFDF), correlation of adjacent frames, prediction residual gradient (PRG) and optical flow gradient (OFG) are extracted from the frames to detect the object based VF.The feature extracted frames are then given to detect the authenticated or forged in the video sequence. To perform this, ResNet152V2 based deep learning model is proposed to detect and localize the object removal VF. Due to its fast training operation, significant learning rates, and our proposed model shows highly efficient outcome in extracting the multi features from the forged video. The ResNet152V2 is a pre-trained model to increase the training process and produces high accuracy rapidly. In addition to this, the ResNet152V2 is incorporated with the bi-directional gated recurrent unit (Bi-GRU) for tuning the hyper parameters and loss function reduction. The improved remora optimization algorithm is introduced to overcome many technical related problems by reducing the complexity of the process. It also maintains the constant performance to optimal the best outcome. This will tends to improve the detection accuracy effectively. Figure (1)manipulates framework of proposed approach.

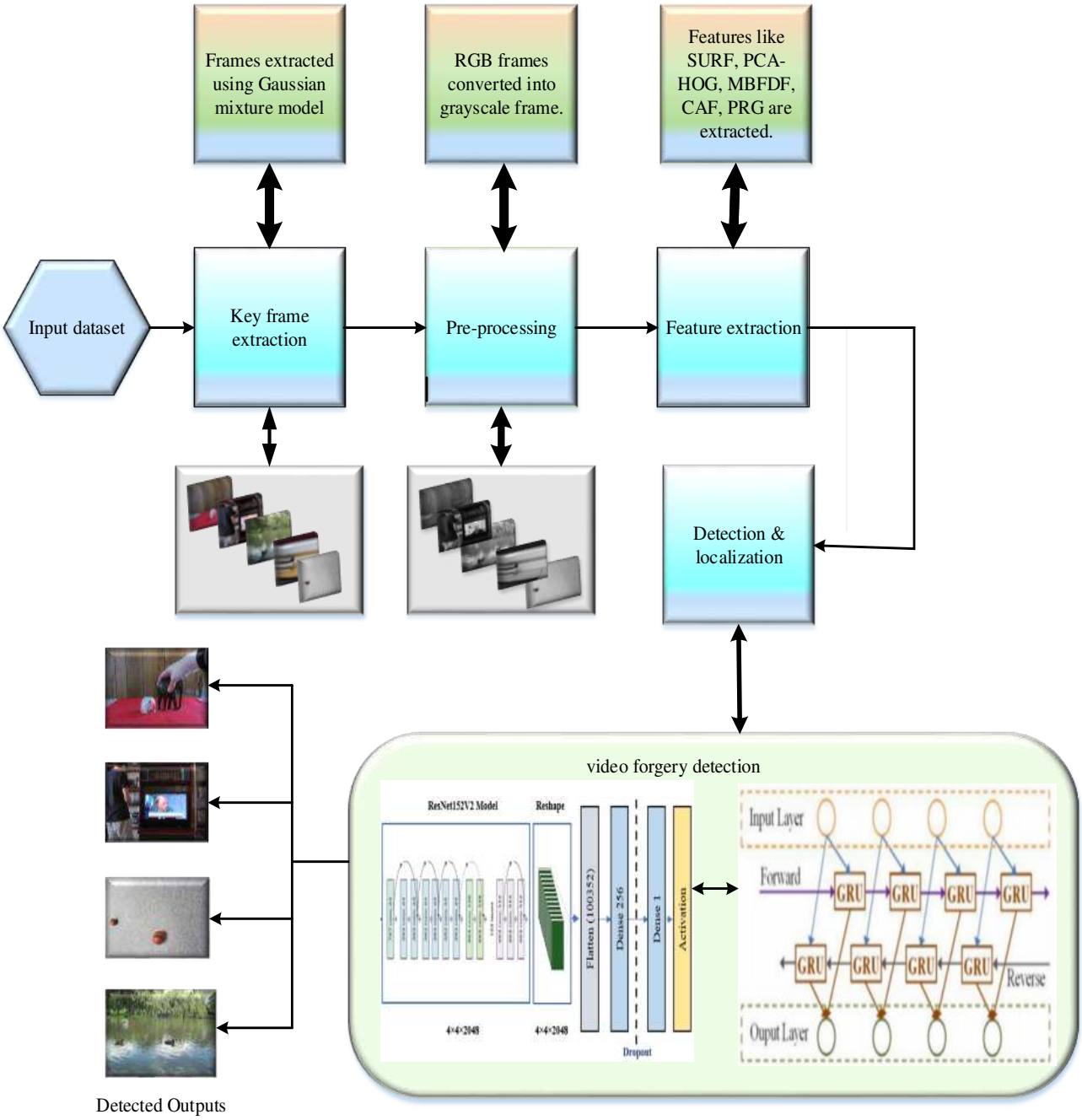


Figure (1) Framework of proposed approach

3.2 Key frame extraction stage

To identify and localize the video forgery, the videos cannot be used directly for implementation. For this, the video is converted into frame for attaining efficient accuracy in the outcome. For this conversion process, a newly evaluated key frame extraction technique Gaussian mixture model (GMM) is emphasized [22]. The GMM technique is the special type of density model that consist of Gaussian function components. For performing the multi-model density, weights of different measures are present in the component of Gaussian function. For separating forefront and background from the source frame sequence, GMM model is highly enhanced.

The value of GMM usually affects the value of background frames. In simple words, when the decreased number of GMM model utilized, the background models of frames are also decreased. Multiple stages are processed for this method, includes stage of equalizing the source to the statistical features and the stage of choosing the statistics that inclined in the background. In GMM, matching stage in each frame, there is a parameter upgrade stage. The expression for GMM model is,

$$p(Y_h)=\sum_{j=1}^{L} w_{jh}\eta(X_h,\mu_{j,t},\sum_{j,h}) \qquad (1)$$

The notation J in (1) is the numeric value of statistics, whereas is the mean value of Gaussian at the unit notation 'h', and covariance matrix is denoted as $\sum$ at J threshold on Gaussian and weight is represented as 'w'. The expressionof Gaussian probability density function.

$$\eta(X_h,\mu_{j,t},\sum_{j,h})=\frac{1}{(2\Pi)^{x/2}\sum^{\frac{1}{2}}}\exp(x_h-\mu_h)^T\sum^{-1(x_h-\mu_h)} \qquad (2)$$

$$w_{j,h}=(1-\alpha)w_{j,h-1}+\alpha B_{(j,h)} \qquad (3)$$

The notation x is the Gaussian distribution size, the value x=1 when the background model is the grayscale frame, $\sum J,H$ is a colour image, and the amount of x=3 is an RGB frame. When the notationB is 1, then it equalizes, and vice versa, the parameter Bis 0. The values of μ and α are upgraded with the equation described below,

$$\mu_{j,h}=(1-\rho)\mu_{j,h-1}+\rho Y_{j,h} \qquad (4)$$

$$\sigma_{j,h}{}^2=(1-\rho)\sigma_{j,h-1}{}^2+\rho(Y_h-\mu_{j,h})^2 \qquad (5)$$

$$\rho=\alpha\eta(Y_h,\mu_{j,h},\sigma_{j,h}) \qquad (6)$$

$$A=\arg\min_a(\sum_{L=1}^{a} W_L>C_g) \qquad (7)$$

Using (3) in the notation upgrade stage, the values of GMM parameters are utilized to run the next source. Upgraded values namely weights, means and variants. The value of weight is upgraded always. After the value of weights get integrated, the entire weight of all distribution is not greater than 1. Then the average value of a statistics is upgraded always there is a particular pixel value that equalizes the statistical value. With (5) and (6) the SD value of statistics is upgraded for regular time interval that equalizes the distribution. Then the equation (7) the pixels with possibility equalizes the Gaussian model, whether they are against the background, they are classified as background frames and on the opposite as the forefront.

3.3 Pre-processing

For detecting the object removal video forgery the color image cannot be processed. Hence, there is a necessary to convert RGB frames into grayscale frame. The numerous amount of consecutive image frames are present in the video sequence. These video frames are constant to each other. But, there is only difference is that among the neighbourhood frames there is a state of moving video objects. The video objects are replicated and moved somewhere in the

sequenced video or destroyed by manifesting or in-painting and it said to be object-based forgery. However, during tampering operation there is a probability of leaving evidence inevitably. For detecting these forgery traces, a subtraction process is provided among each frames. This highly helps to minimize the redundancy and then cut the residual sequence to residual frame patches using subtraction operation. In pre-processing stage different kinds of noises can be reduced more efficiently.

However, to input the video frames into DNN model, the video frames are transformed into motion residual frames. The source video sequence of the transformed video frames of length M as,

$$B=\{K_1, K_2,...K_j..K_M\}\ j\in\{1,....M\} \tag{8}$$

Here, K_j indicates the *jth* transformed video frame.

Moreover, the transformed video frame is expanded from the advanced video with advanced encoding process, in addition to this, the RGB with *R, G, B*components. To get rid of high computational complexity, the transformed video frame is transformed into grayscale frames and then undergoes subtraction operation. Finally, the obtained gray scale frame is denoted as,

$$S_t=(gray(H_t)-gray(H_{t-1}))\ t\in\{2,.....,M\} \tag{9}$$

Here the function of *gray* () denotes the color space is converting RGB colour into gray-scale. The subtraction starts from the video of second frame. Hence, the resultant S_t can be indicated as 8-bit grayscale frame that demonstrates the motion residual between following video frames.

3.4 Multiple feature extraction approach

This approach is very important for extracting different features from each frames. Here, the pre-processed frames are then given to multiple features such as improving robust features (SURF), (PCA-HOG) feature, (MBFDF), CAF, (PRG) and (OFG) based features [21] are extracted. By extracting these features, can study the nature of the video in efficient manner. Through this approach object-based VF can be detected and localized more accurately. The upcoming section describes the different multi-features in detail below:

3.4.1 SURF based features

To improve the stability and accuracy of the detection SURF process is highly suitable and compare the frames in a circumstantial, equally invariant manner. The key operation of the SURF method is that it has the capability for computing the operators instantly with the aid of box filters. These methods are used in many practical applications such as surveillance and scene understanding. It undergoes two stages namely feature withdrawal and feature designation. In our research work, the SURF features are withdraw from J^{pre} .

- **Feature withdrawal**

  The feature extraction aids to detect the object using simple hessian matrix approximation. For performing hessian matrix, initially convolution is enhanced with Gaussian kernel. Then, the second order derivative called scalar invariant fourier

transform (SIFT) to coordinate with LoG approximation. The SURF method utilizes box filters to enhance the simplification for both convolution and second derivative.

- **Feature designation**

  The SURF designator undergoes two major process: (i) redeveloping orientation fixation in a ring region with the details on the key point and (ii) calibration of the choice based places for orientation and SURF designator withdraw from it. V^{SURF} denotes the extracted features of SURF.

3.4.2 PCA-HOG features

HOG: HOG feature is determined as a feature designator for detecting J^{pre} of same field. There are four stages involved in HOG feature extraction and it is described below:

- ✓ Separate J^{pre} on the basis of smaller interlinked cells and calibrate a histogram of slope locations or edge inclination for every frame present in the cells.
- ✓ Each cell is inclined as angular bins with the aid of gradient orientation.
- ✓ Each frame is presented with weighted slope with respect to the angular bin.
- ✓ In the dimensional section, each group of neighbouring blocks is considered. This group of histogram can be emphasized using block histogram, and these groups are denoted as designator. The categorization of cells in each blocks is determined by grouping and normalization of histograms. The extracted HOG features are denoted by T^{HOG}

PCA: PCA is used to determine the process of determining the principal components and utilized to perform a variation on the basis of data, may be few components are utilized and other elements are omitted. The mathematical expressions of features are described in the oncoming division.

Standard deviation (SD)

The mean variability among the mean and the point at which the information is calibrated with the rid of squaring them and it is denoted by SD. The mathematical expression is provided in equation (10).

$$SD = \sqrt{\frac{1}{p}\sum_{e=1}^{p}(V_e - \overline{V})^2} \tag{10}$$

Here, V indicates the random number and p denotes the sample size.

Covariance

The weight of variation from the average is designated as covariance. The mathematical expression for covariance is manifested as equation (11).

$$Cov(V,J) = \frac{\sum_{e=1}^{p}(V_e - \overline{V})(V_e - \overline{V})}{p} \tag{11}$$

Mean

The average of the data in asummarized variable V as $\overline{V} = \frac{1}{p}\sum_{e=1}^{x} V_t$. The extraction of PCA-HOG feature is demonstrated as, $T^{PCA-HOG}$

3.4.3 MBFDF

The MBFDF mainly aims to show the variation to SC frame from the equalized DC frame. ϑ^2 represents divergence and it is manipulated in equation(12) and this feature is used to demonstrate the stability of fixing. In addition, the first order distribution of each coefficient is evaluated as notation $Q_c(m)$, and in the c^{th} mode, the theoretical distribution is presented as $\hat{Q}_c(m)$. The MBFDF feature extraction is indicated as T^{MBFDF}.

$$\vartheta^2 = \sum_{v=1}^{x} \frac{(Q_c(m) - \hat{Q}_c(m))^2}{\hat{Q}_c(m)} \tag{12}$$

3.4.4 Correlation of adjacent frames

The equalization of nearest frames is utilized to calibrate the summary accordingto the inter-frame content. Among $c^{th}\ and\ (c+1)^{th}$ frame the correlation coefficient is denoted as f_c. The two dimensional PC value of c^{th} frame at (U, V) area is represented as $G_c(U,V)$.

$$f_l = \frac{\sum_x \sum_y (D_c(U,V) - \overline{D}_b).J_{c+1}(U,V) - \overline{D}_{c+1})}{\sqrt{(D_c(U,V) - \overline{D}_c)^2.(D_{c+1}(U,V) - \overline{D}_{c+1})^2}} \tag{13}$$

Moreover, the c = 1, 2,….., s-1 and D denotes the phase congruency. In addition to this, s indicates the total count of frames and $\overline{D}_b$ indicates the average two dimensional PC for c^{th} frame. The expressionis mentioned in the equation (14).

$$D_c = \frac{1}{t \times q}\sum_{x,y} D_c(U,V) \tag{14}$$

Here, t indicates thickness of the frames in pixels and q denotes the height of the video. The feature extracted correlation coefficient is represented as T^{CAF}.

3.4.5 PRG

For the detection of video forgeries, the prediction residual principle is high emphasized. The error among the preceeding and succeding frame of the original frame is said to be as residual prediction. In addition with this, information about the similarity of the adjacent frames. However, the uniform cell-equalizing method is manipulated for identifying the frames which are present in the future. It provides the relationship in each frames with the help of information provided in the reference frame of adjacent cells. The residual calculation is the comparison of MSE for every pixel size of 16×16 to alternate the coequal in the reference frame.

$$pr(j) = frame(j+1) - LC(frame(j)) \quad (15)$$

$$prg_j = pr_j + 1 - pr_j \quad (16)$$

In equation (15), frame(j) denotes the frames and the corresponding frames are predicted by the cell equalization method which is represented as LC(..). Moreover, for frame(j+1), probability is calculated with the help of LC(frame(j)). After that, j^{th} pair frame, the prediction residual (*pr*) is measured with the help of computing the varaibility among the frame (j+1) and the frame (j+1) is predicted by LC(frame(j)). The PRG feature extraction is denoted by T^{PRG}.

3.4.6 OFG

The optical flow is the performance of the moving illumination pattern between the nearby pixels. It is also used in monitoring the illumination difference from one x frame to the nearby frame of the pixel. During the time t, the illumination of the pixel t (u, v) on the basis of the frame be $EHS(u, v\,t)$. The OFG feature extraction is mathematically expressed in equation (17) and (18), respectively.

$$Oflow_j = \iint \{(E_u g + E_v h + E_t)^2 + \beta^2[(\|\Delta g\|)^2 + (\|\Delta x\|)^2]\} dxdy \quad (17)$$

$$Oflowg_j = Oflow_{j+1} - Oflow_j \quad (18)$$

Where, the notation E_u, E_v, E_t denotes the derivatives corresponding to the strength along with the direction u, v, t and time width, accordingly. The OFG feature can be manipulated as T^{PRG}.

The combination of all extracted features are represented in mathematical expression as given below,

$$T = T^{OFG} + T^{PRG} + T^{CAF} + T^{MBFDF} + T^{PCA-HOF} + T^{SURF} \quad (19)$$

3.5 Forgery detection and localization using ResNet152V2 with Bi-GRU

Residual network or ResNet is one of the convolutional neural network. The input image size of $240 \times 340 \times 3$ forged frame is processed. This model consist of high initial weights because of its pre-trained model can help to gain efficient accuracy than the traditional CNN. ResNet152 version 2 based ResNet consist of 152 layers followed by reshape layer, flatten layer, a dense layer with forged frames, a dropout layer and additionally a dense layer with softmax activation function to display the forged and authenticated frame. Due to the rise of layers in deep network, the weight of the layers cannot be increased accurately between the layers. Figure (2) represents schematic diagram of Resnet152V2 model with Bi-GRU

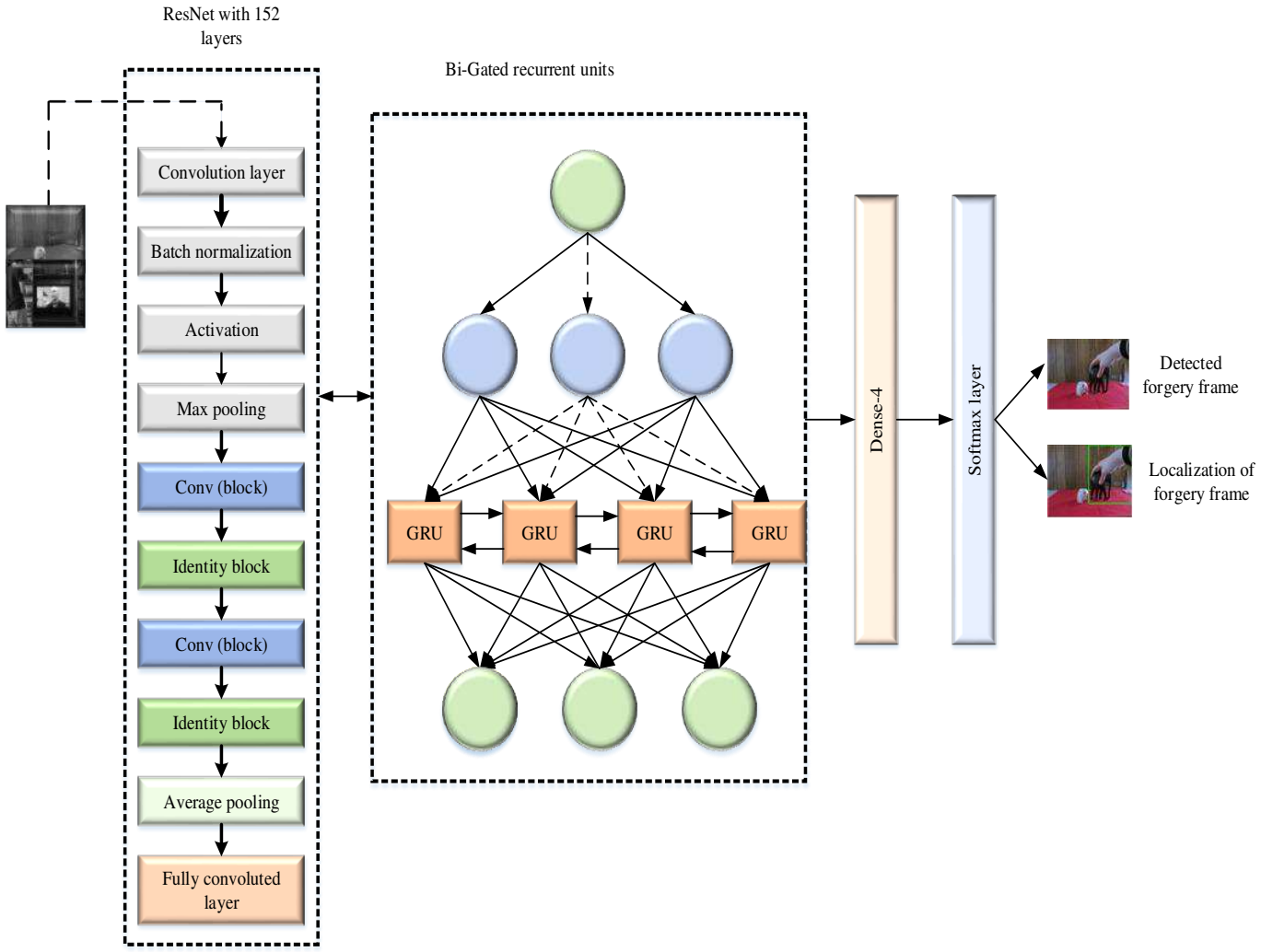


Figure (2): Schematic diagram of Resnet152V2 model with Bi-GRU

To overcome this degradation issue, ResNet152 uses tiny link on the parallel via ordinary convolution layers. The outcome of ResNet is represented mathematically as,

$$D(x) = E(x) + x \tag{20}$$

Where, $D(x)$ represents the residual block output

The stocked weight layer non-linearity is represented as,

$$E(x) = D(x) - x \tag{21}$$

Where, $E(x)$ represents the stocked weight layer non-linear.

In addition to this, gated recurrent unit (GRU) is introduced and it is one of the recurrent neural network (RNN) architecture. The main advantage of GRU is it has the ability to maintain information peripherally with the prediction for a long time without deleting it.It is very easy to modify and takes only less training time that makes the accuracy more efficiently. In our proposed method, bi-directional gated recurrent unit (Bi-GRU) is

introduced. The Bi-GRU is used to overcome the sequential complexity of the ResNet model. The single GRU at time step p is defined mathematically as,

$$u_p = \alpha(A^u y_p + B^u J_{s-1}) \qquad (22)$$

$$f_j = \alpha(A^f y_j + BfuJ_{s-1}) \qquad (23)$$

$$J_j = (1-u_p)\Theta\ J_{j-1} + u_s \Theta d_j) \qquad (24)$$

$$d_j = w(A^J y_j + B^J (f_j \Theta J_{j-1})) \qquad (25)$$

Here, y_j is the element in input sequence; d_j indicates the outcome measured; u_p represents the upgraded gate; f_j denotes the forget gate; $\alpha, w and \Theta$ each presents the activation function sigmoid; J_j denotes the concealed gate *tanh* activation function and component multiplication. In the bi-directional gated recurrent unit, the hidden state of forward and backward travelling are joined as the following reserved representation $J^{\lambda}{}_j$:

$$J^{\lambda}{}_j = [GR\overset{\cup}{U}(\sigma_j); GR\overset{\cup}{U}(\sigma_j)] \qquad (26)$$

Since, the equation for σ_j with the aid of increment operation of the weight Y-th portion presentation and the token manipulated by dividing $t = \sum_{y=1}^{Y} 2^y$ is formulated mathematically as,

$$\sigma_j = \sum_{y=1}^{Y} \frac{b_y J^y{}_j}{t} \qquad (27)$$

Here, b_y represents the weights Y-th layer representation. Then, $b_y = 2^j$ is chosen in this research.

The bi-GRU is the series of proceeding model, one using source in the forward direction and other in the backward direction. Bi-GRU works in parallel, with the rid of parallel computation that can perform multiple calculation at the same time. One way GRUs cannot control multiple operations at the same time. The Bi-GRU considers the previous sequence and succeeding sequence of the frames, to avoid the noise during processing. This model shows faster performance with low time duration with efficient extended operation.The Resnet152V2 followed by Bi-GRU used to detect and localize the object based VF.It consist of drop out layer with a softmax activation function to detect and localize the videos based on forgery detection and forgery localization. However, the accuracy of this approach is highly affected due to loss function and hyper parameters. To overcome this issue, improved remora optimization algorithm is emphasized. The improved version of remora algorithm helps to fine tune the hyper parameters and loss function reduction. The loss function reduction and tuning of hyper parameters using optimization algorithm is described below.

Loss function

The detection and localization of forgery videos are highly disturbed due to loss function. Reduction of loss function in DNN makes the system work efficiently with the multiple features. The loss function always focus on the negative frames and hence leads to low true

positive rate. During training process, there will exist a small imbalance to handle the entire network. This imbalance assigns the extra factor to the original cross degenerated term. So, the loss have the ability to overcome the gradient of different imbalance frames. The mathematical expression for loss function is formulated below,

$$L(x,z) = -\sum \beta(1-z)^{\lambda} * x \log(z) - \sum (1-\beta) z^{\lambda} * (1-x) \log(1-z) \tag{28}$$

Here, x indicates the binary ground truth mask and z denotes the prediction pixel. λ and β are hyper parameters.

Loss function reduction using ROA

To upgrade the weight standards and to enhance the reliability of the detection and localization, this research paper manipulated the remora optimization approach. This algorithm is on the rid of remora, the suckerfish from the fish family of Echeneidae. This approach help to compute the world-wide search of images with enhanced grayscale values. It major aim is to reduce the multi-thresholding time complexity and it can eliminate the region of interest being emphasized using local trapping.

The remora optimization algorithm mainly comes under the concept of remora whale and it is considered as expertised traveler in the ocean. There are two stages involved in this algorithm they are, exploration and exploitation. The nature of remora is to travel free and peaceful eating and it can be presented in mathematical expression. Here, mode converting and precision of this optimization can be enhanced through remora factor that increases the convergence. For mode conversion decisions, the stages like travel free, eating peacefully etc. these methodologies aids the RO algorithm to obtain optimal outcome. Various steps involved below and are described in detailed manner:

(a) Primary approach

The best problem solver called remora, and its present location V is the movement, that evaluates the problems in the search place. Remora whales location changes based on the size of the pool. The present position is manipulated as,

$$V_j = (V_{j1}, V_{j2}, ..., V_{jv}) \tag{29}$$

In final words, each problem solving evaluates its own fitness value. The fitness function can be expressed mathematically as,

$$fitness = \min(L(x,z)) \tag{30}$$

- Travel free (Exploration)

The location of the remora can be upgraded when it is etched with swordfish and it can be mathematically formulated as,

$$V_j^{h+1} = V_{opt}^{h} [rand(0,1) * (\frac{V_{opt}^{h} + V_{rand}^{h}}{2} - V_{rand}^{h}] \tag{31}$$

Where, V_j^{h+1} is the present location of the remora with its number (j) and H denotes the extended number of iterations and h indicates the ongoing recursions. The continuous changing of remora is manipulated as $V_{rand}\ and V_{opt}^{h}$ is the best location of the remora.

In addition to this, remora may alter the host based on its experience. It is mathematically expressed as,

$$V_j'(h+1)=V_j(h+1)+randn\times(V_j(h+1)-G_j(h)) \tag{32}$$

- Eat thoughtfully (Exploitation)

Remora also can attach themselves with the humpback whales for food. Hence the remora have the flexible characteristics of humpback whales. The whale optimization algorithm (WOA) is manipulated with ROA to implement the local search. Moreover, the bubble-net attacking method used in WOA is enhanced. The mathematical expression for position upgrading stage is as follows:

$$V_j(h+1)=G\times e^b\times\cos(2\prod b)+Y_{best}(h) \tag{33}$$

$$G=\left|Y_{best}(h)-Y_j(h)\right| \tag{34}$$

$$b=rand\times(c-1)+1 \tag{35}$$

$$c=-(1+\frac{h}{H}) \tag{36}$$

Here, G indicates the distance among remora and food. According to equation (35) and (36), it is known that b represents the random number between -2 and 1. Here, c reduces linearly from -1 and -2.

Improved remora optimization algorithm using AFM

The ROA is mainly introduced based on parasitic feeding of whales and swordfish. In simple words, remora can search its own food. In accordance with this, a new autonomous foraging mechanism is introduced in the ordinary remora optimization algorithm to provide high intensity in searching food. The improved remora optimization is more flexible and maintain a good stability among exploration and exploitation compared to ordinary approach. The basic ROA shows high computational complexity with low accuracy. The autonomous foraging mechanism (AFM) is described below:

- Autonomous foraging mechanism (AFM)

  In basic ROA, the food is find randomly and produce the food on the basis of present food position. But in improved ROA with AFM, two different operators are enhanced to expand the optimization capability of ROA. Initially, the remora has the chance z of seeking for food in unknown locations. When, remora usually searches its space and food widely and randomly. The formula for AFM is,

$$Y_j(h+1)=(UE-LE)\times rand+LE \tag{37}$$

Here, the UE and LE indicates the upper edge and lower edge of the search space respectively.

Based on equation (37), the initial operator belongs to the exploration capability of the ROA, preventing the local optimal values effectively. In other words, to increase the exploration capability of ROA, the second operator is given to division operator and multiplication operator and the location upgrading equations are defined as follows,

$$Y_j(h+1)=\begin{cases} Ybest(h)\div(RMOP+\xi)\times((UE-LE)\times\lambda+LE)\times levy, & rand<0.5 \\ Ybest(h)\times RMOP\times((UE-LE)\times\lambda+LE)\times levy, & rand\geq 0.5 \end{cases} \quad (38)$$

$$RMOP=1-(\frac{h}{H})^{1/\beta} \quad (39)$$

$$\beta=10\times rand-1 \quad (40)$$

Here, the RMOP denotes the random math optimizer probability, measured by the present number of recursions, increased number of recursions, and parameter β. Based on equation (40), β indicates the random number among -1 and 9.

In equation (38), the changed position is developed in accordance with the present best location. Here, levy operator is proposed to maximize the population diversity. Each remora conducts the second operator when rand < d, where d indicates the parameter.

5. Results and discussion:

This division represents the outcome of the proposed methodology to show the operation efficiency in comparison with other existing approaches. The work has been manipulated with the implementation tool python. The efficiency of the proposed method is emphasized via VF dataset. In this work, for detection and localization of VF we take the sample videos from REWIND dataset and it is manipulated in the GUI interfaced. This dataset consist of total amount of 80 video sequence of about 320×240 pixel resolution and 30 fps frame-rate. The video sequence has been constructed initially from 10 original video sequence, that has been saved by low-end devices and narrow at the origin (by either MJPEG or H264 on the utilized device). In each video sequence, the forged video has been evaluated via copy-move. Forged videos have been recorded by lossless coding to eliminate developing any additional artifacts such as noise, blurring, rotation etc. hence, the forged video sequence is different from original video only on tampered pixels. Total original and forged video sequence have been orderly compressed using H264 with constant QP= 10, 20, 30, and additionally 60 videos are obtained. Here, each forged sequence is provided with a MAT file, that consist of mask indicates which pixels have been remodeled by the forgery (considering R component only in the RGB color space). This results as base truth to detect and localize forgery in the video sequence. The discrimination and the performance analysis are estimated in the below sections. Table 2 tabulates the hyper parameter of the proposed work and Table 3 tabulates the device configuration of the developed model.

Serial No.	**Hyper parameter**	**ResNet152V2 –Bi-GRU with IROA**
1	Learning approach	Remora

2	learning rate	0.06
3	Batch size	64
4	Epoch	100
5	Input size	(240*320*3)

Table 2: Hyper-parameter setting of the proposed work

Serial No.	Parameters	Configuration
1	Device name	ssm113.smg.local
2	Processor	Intel (R) Core (TM) i7-8700 CPU @ 3.20GHz, 3.19GHz
3	Installed RAM	8.00 GB (7.85 GB unstable)
4	Device ID	303461EA-28F3-4C81-92BD-DB467FA392F8
5	Product ID	00331-60000-00000-AA830
6	System type	64-bit operating system, ×64-based processor
7	Pen and touch	No pen or touch input is available for this display

Table 3: Device configuration of the developed model

5.1 Performance metrics

Accuracy:

In the performance analysis, the accuracy is one of the most significant measure to evaluate the proposed method efficiency and enhancement rate. The accuracy predicts the correct solution from the number of cases examined. To compute the accuracy by considering the following expression:

$$A_y = \frac{Tp + Tn}{Tp + Tn + Fp + Fn} \tag{41}$$

where, the true positive is Tp, the true negative is Tn , the false positive is represented as, Fp and the false negative is denoted as, Fn.

Precision:

The precision is also one of the efficient performance analysis to predict the exactness of the proposed method. It is the number of information which is transferred by a value. The precision mathematical formula is expressed below:

$$P_n = \frac{Tp}{Tp + Fp} \tag{42}$$

where, Tp is the true positive value and Fp is the false positive value.

Specificity:

Specificity is another type of performance metrics to find out the image get forged or not in flawlessly. It tests the true negative value. The expression for specificity is given as follows:

$$TNR = \frac{TN}{TN + FP} \tag{43}$$

Where, Tn denotes the true negative value.

Sensitivity:

Sensitivity is also called as recall and it is termed as the smallest amount of changes that is appeared in the videos can be detected accurately. It is an absolute quantity. It states that the ratio of number of true positives to the combination of number of true positives and the false positives. Mathematically the sensitivity is derived as,

$$Sensitivity = \frac{TP}{TP + FN} \tag{44}$$

F-measure:

The F-measure states that it the comparative measures of precision and recall and the harmonic mean of precision and recall. The precision and recall is said to be maximum and perfect if the F1-score returns the values as 1. If the value of recall or precision is zero means the F1-score becomes low i.e. zero. The mathematical formulation of F-measure is expressed below:

$$F - measure = \frac{2 \times \mathrm{Pr}ecision \times Sensitivity(\mathrm{Re}call)}{\mathrm{Pr}ecision + Sensitivity} \tag{45}$$

False positive rate (FPR):

The FPR is otherwise called as false alarm ratio or fall out. If the value of FP is zero then it provides better result i.e. no false positive. It is the ratio of invaluable negative events that are mistakenly predicted as positive to the entire number of negative events. The FPR is derived by using the following equation:

$$FPR = \frac{Fp}{Fp + Tn} \tag{46}$$

False negative rate (FNR):

FNR is also termed as miss rate. When the FP is zero the FP rate is also becomes zero i.e. no false negative. The FNR is the division in which the total amount of false negatives by the amount of true positives. The FNR is mathematically formulated as follows,

$$FNR = \frac{Fn}{Fn+Tp} \tag{47}$$

Matthew's correlation coefficient (MCC):

MCC states that the correlation between the true values and the values which are predicted. It is same as that of Pearson's correlation and ranges from -1 to 1. When its returns the value as 1.0 means it detect perfectly otherwise the detection is imperfect. The expression of MCC is given in mathematical formulation as,

$$MCC = \frac{TP*TN - FP*FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{48}$$

Negative predictive value (NPV):

NPV is defined as, in a perfect detection if it returns no false negative means the NPV becomes 1 i.e. it attains maximum. Otherwise the value of NPV is zero because it gives no true negative. The NPV formula is stated as follow,

$$NPV = \frac{TN}{TN+FN} \tag{49}$$

False discovery rate (FDR):

FDR is determined as the ratio of amount of false positive detection to the total amount of false positive and true positive detections. It is expressed as follow,

$$FDR = \frac{FP}{(FP+TP)} \tag{50}$$

Mean squared error (MSE):

MSE is defined as the subtraction of square of difference between the actual and predicted values to the total number of actuals values. The mathematical model of MSE is expressed below,

$$MSE = \frac{\Sigma(y_i - \hat{y}_i)^2}{n} \tag{51}$$

Here, y_i represents the actual value, $\hat{y}_i$ is the predicted value and n denotes the total number of actual values.

Peak signal-to-noise ratio (PSNR):

The PSNR is defined as the ratio between the noise which affects the video and the maximum power signals. The reconstruction quality is higher if the value of PSNR is maximum. The formula for PSNR is given below,

$$PSNR = 10\log_{10}\left(\frac{(\max(G_i))^2}{MSE}\right) \tag{52}$$

Where, G_i represents the grey level intensity values.

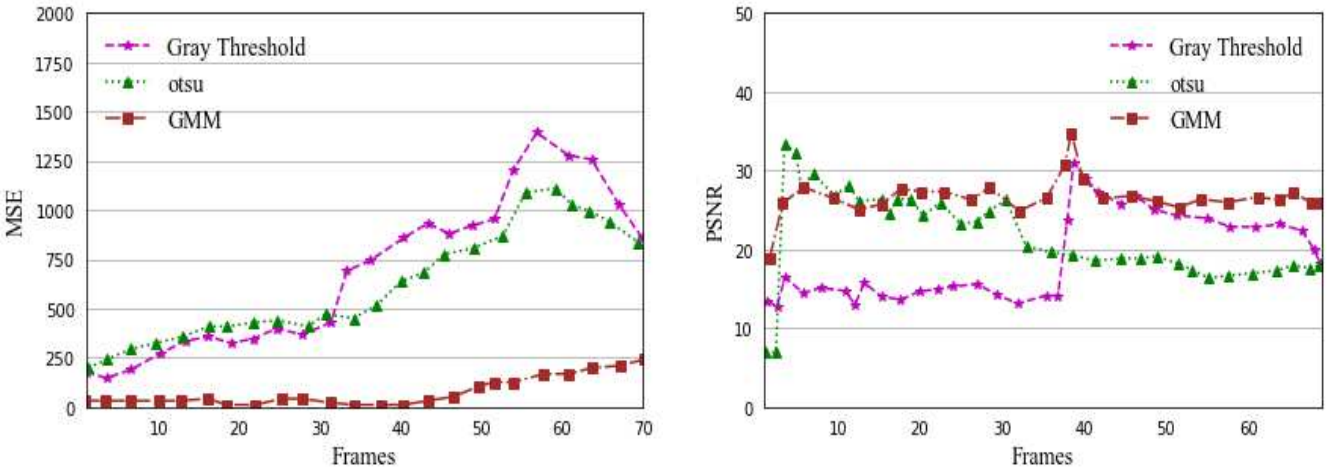
5.2 Analysis of VFD and localization systems

The major aim of our research is VFD and localization. For this our research work undergoes four stages namely, key frame extraction, pre-processing of noisy frames, feature extraction and detection. Initially, the videos are converted into frames using GMM model and then given to pre-processing stage. In pre-processing step, the RGB frame is converted into grayscale image. From the grayscale frame multiple features such as SURF, PCA-HOG, MBFDF, CAF and PRG are extracted. The feature extracted frames are then utilized for detection and localization of forged videos. The upcoming section provides the detailed explanation about the analysis of our proposed model.

5.2.1 Key frame extraction stage

In this step, the frames are extracted from the videos to process the entire research work. For this, Gaussian mixture model approach is emphasized to extract the frames from the video sequence. The GMM model mostly consider the density of the frame and compares with the Gaussian function components. This model is highly convenient for separating forefront and background from the input video sequence. The evaluated performance was MSE and PSNR and it is plotted in graph. The frame extracted from the video sequence is displayed in Figure (3)

In the VFD techniques, the Figure (a) and (b) shows the adaptive threshold of MSE and PSNR. The figure (4) shows the best result for MSE and PSNR values i.e. the achieved outcomes for MSE is 104.02 and the PSNR is 27. These values are more efficiency when distinguishedwith the existing Gray threshold and Otsu approaches. The gray threshold computation complexity is high so it is not good for the real-time process and it is very sensitive for its window size, its noise level and its parameters. In Otsu the major drawback which is faced here is Gaussian noise due to this noise the quality of video get decreased. The MSE values of existing forgery detection methods are: the gray threshold is 645.38 and the Otsu method is 595.36. The PSNR values of existing gray threshold and Otsu are: 19.36 and 20.66 respectively. So we go for the proposed GMM model it uses some pre-trained model to speed up the processing time and produces high quality videos also. The achieved values of MSE and PSNR clearly shows that the proposed model produces high efficient and maximum quality videos.



(a) MSE (b) PSNR

Figure (3) Analysis of **(a)** MSE and **(b)** PSNR.

Figure (4) key-frames extracted by GMM approach

5.2.2 Pre-processing and multiple feature extraction

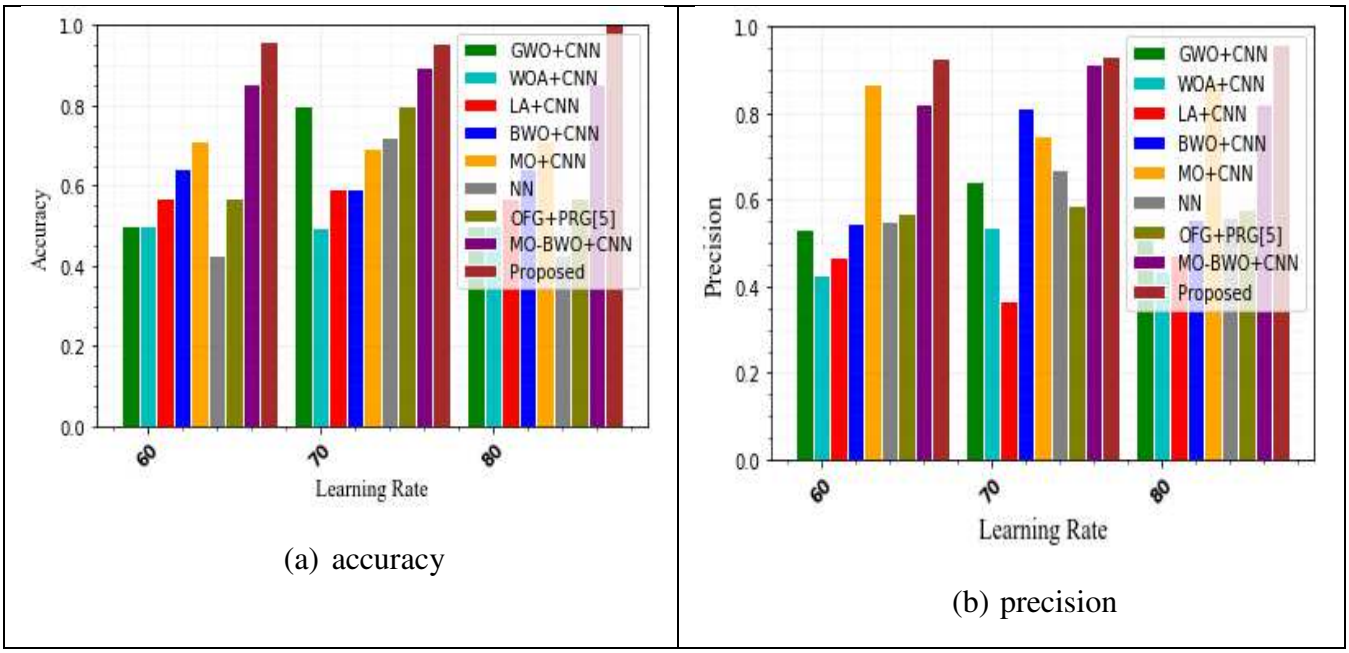
In preprocessing stage, RGB color frame is converted into grayscale frame. However, this stage also removes unwanted noise such as temporal noise, Gaussian noise to attain better accuracy and efficiency. The pre-processed frames are then given to extract multiple features from the frame of the video sequence. The features which are extracted from the video frames are: SURF, PCA-HOG, MBFDF, CAF and PRG. The pre-processed frames are presented in figure (5).

Figure (5) Pre-processed frames

5.2.3 Detection and localization using DL approach

The video forgery can be detected and localized with the aid of newly approached deep learning model ResNet152V2 with Bi-GRU. It consist of drop out layer with a softmax activation function to detect and localize the videos based on forgery detection and forgery localization

Figure 6(a) illustrates the accuracy measure with different learning rate like 60, 70 and 80 correspondingly. When the learning rate is 70, the achieved accuracy values of proposed GMM with the existing approaches such as, GWO+CNN, WOA+CNN, LA+CNN, BWO+CNN, MO+CNN, NN, OFG+PRG [5], MO-BWO+CNN are: 96%, 75%, 53%, 60%, 60%, 75%, 78%, 80% and 85% respectively. The Figure 7(b) shows the precision values of proposed and existing approaches when compared the proposed methodology to the previous methods the attained precision value of the proposed GMM is maximum i.e. 96%. If we considered the learning value 70 to prove the proposed achieved maximum precision than existing which are: 62%, 55%, 35%, 80%, 75%, 64%, 60%, and 89% respectively.



(a) accuracy

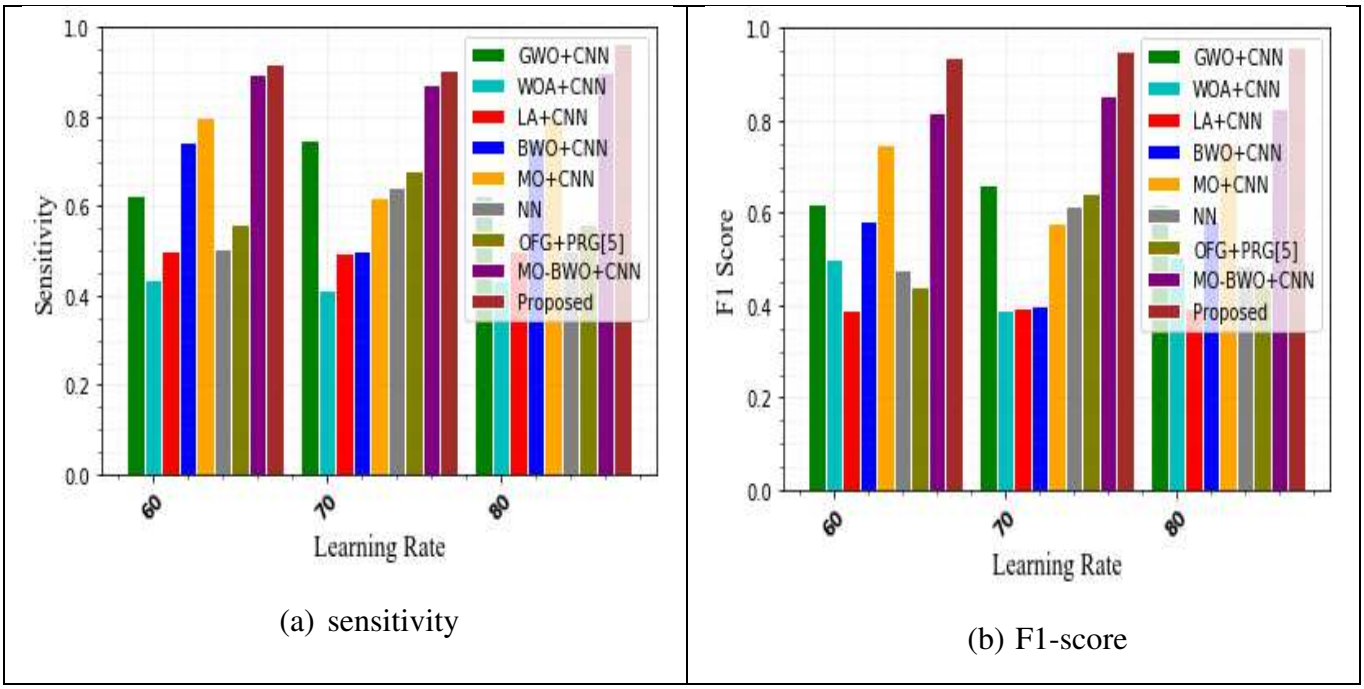
(b) precision

Figure (6) Performance analysis of **(a)** Accuracy and **(b)** Precision with various learning rate.

Methodologies	**Learning rate 70 (%)**		**Learning rate 80 (%)**	
	Accuracy (%)	**Precision (%)**	**Accuracy (%)**	**Precision (%)**
GWO+CNN	75	62	42	55
WOA+CNN	53	55	42	40
LA+CNN	60	35	58	42
BWO+CNN	60	80	62	56
MO+CNN	75	75	70	90
NN	78	64	40	56
OFG+PRG [5]	80	60	56	80
MO-BWO+CNN	85	89	84	58
Proposed	**95**	**96**	**99**	**98**

Table 4: Comparison table for accuracy and precision for the learning rate 70 and 80.

Table 4 Comparison table for accuracy and precision for the learning rate 70 and 80. Figure 7(a) and 8(b) illustrates the performance efficiency of sensitivity and F1-score by using different training sizes. For the sensitivity measure, the values achieved in the proposed and the existing method such as, GWO+CNN, WOA+CNN, LA+CNN, BWO+CNN, MO+CNN, NN, OFG+PRG [5], MO-BWO+CNN in the learning rate 70 are:96%, 75%, 40%, 55%, 57%, 60%, 65%, 70%, 85% respectively. These sensitivity values clearly shows that the proposed GMM accomplishes maximum sensitivity values than other approaches. Figure 8(b) gives the clear explanation for the F1-measure achievements, it gains the increase F1-score values of 96% for the proposed GMM than the existing techniques: GWO+CNN, WOA+CNN, LA+CNN, BWO+CNN, MO+CNN, NN, OFG+PRG [5], MO-BWO+CNN F1-score achieved rates when the learning rate is 70 are: 62%, 37%, 38%, 40%, 58%, 60%, 62%, 83% respectively.



(a) sensitivity

(b) F1-score

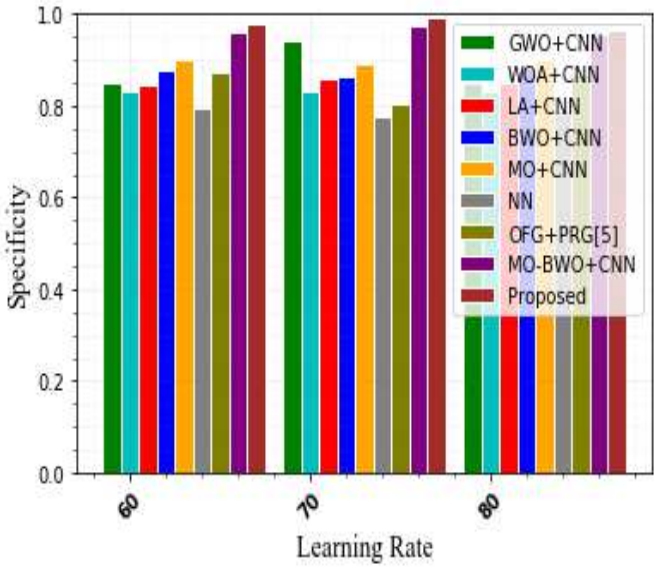
Figure (7) Performance metrics of **(a)** Sensitivity and **(b)** F1-Score with different learning rate.

Methodologies	**Learning rate 70 (%)**		**Learning rate 80 (%)**	
	Sensitivity (%)	**F1-score (%)**	**Sensitivity (%)**	**F1-score (%)**
GWO+CNN	75	62	62	62
WOA+CNN	40	37	40	50
LA+CNN	55	38	50	38
BWO+CNN	57	40	75	58
MO+CNN	60	58	80	75
NN	65	60	50	48
OFG+PRG [5]	70	62	55	43

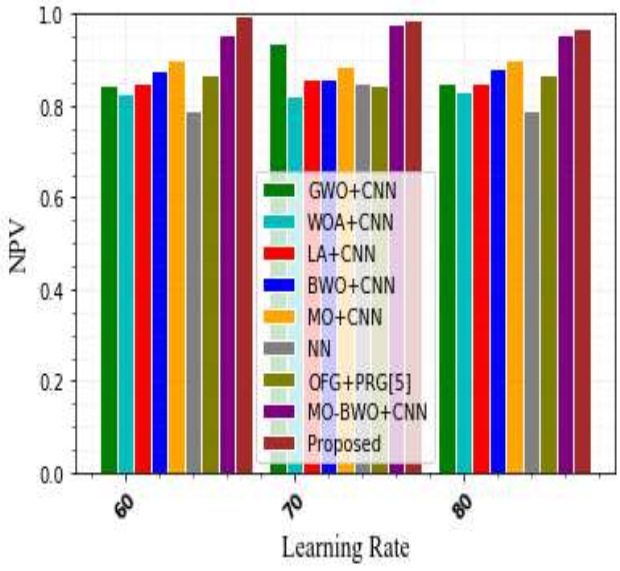
MO-BWO+CNN	85	83	85	80
Proposed	**90**	**93**	**98**	**98.7**

Table 5: Comparison table for Sensitivity and F1-score by varying learning rate 70 and 80

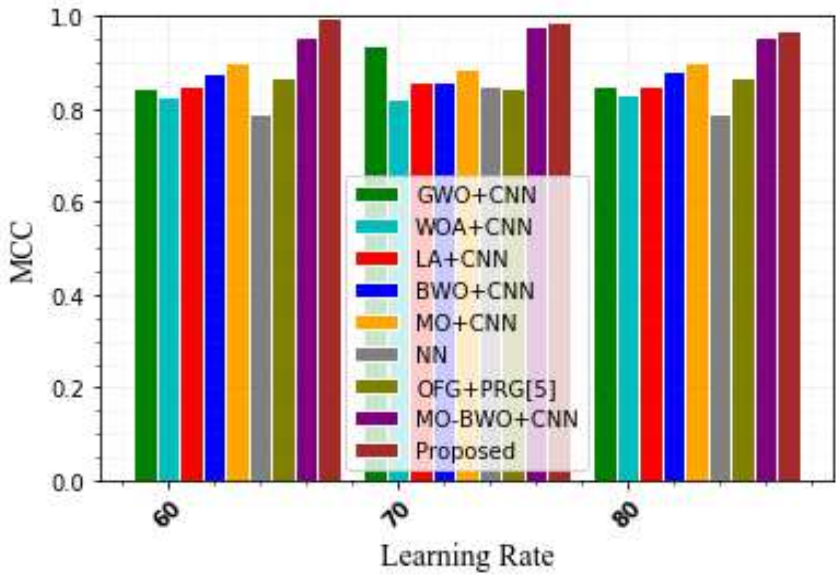
Table 5 Comparison table for Sensitivity and F1-score by varying learning rate 70 and 80. The specificity measure of the proposed technique and some of the existing approaches such as, GWO+CNN, WOA+CNN, LA+CNN, BWO+CNN, MO+CNN, NN, OFG+PRG [5], MO-BWO+CNN are clearly appears in the Figure 8(a). If the learning rate is 70, the gained results of proposed method and the existing approaches are: 96%, 88%, 82%, 85%, 86%, 87%, 76%, 80%, 93% with the help of these specificity values the proposed GMM specificity performance is higher than the previous video forgery detection methods. Figure 8(b) is the performance measure of NPV by varying learning rates as 60, 70 and 80. In that, the various existing video forgery detection techniques are compared with the proposed to prove the proposed GMM accomplishes maximum NPV outcome than the existing systems. If considering the values of proposed and existing techniques in the learning rate 70, the proposed GMM is 6% enhanced than the GWO+CNN, 14% enhanced than the WOA+CNN, 17% increased than the LA+CNN, 17% increased than the BWO+CNN, 8% enhanced than the MO+CNN, 17% improved than NN, 17% improved than the OFG+PRG [5] and 3% enhanced than MO-BWO+CNN. It clearly indicates the proposed GMM method achieves 96% of maximum NPV value than the other existing approaches. The MCC performance analysis is given in the figure 8(c), it illustrates the comparison of proposed GMM and the various existing video forgery detection techniques. In the learning rate 70, the attained MCC value of proposed technique is 92.5 i.e. the proposed GMM is 20% superior than the implemented GWO+CNN, 33% superior than existing WOA+CNN, 34% superior than LA+CNN, 30% superior than BWO+CNN, 47% superior than MO+CNN, 36% superior than NN, 27% superior than OFG+PRG [5], 10% superior than MO-BWO+CNN respectively.



(a) specificity



(b) NPV



(c) MCC

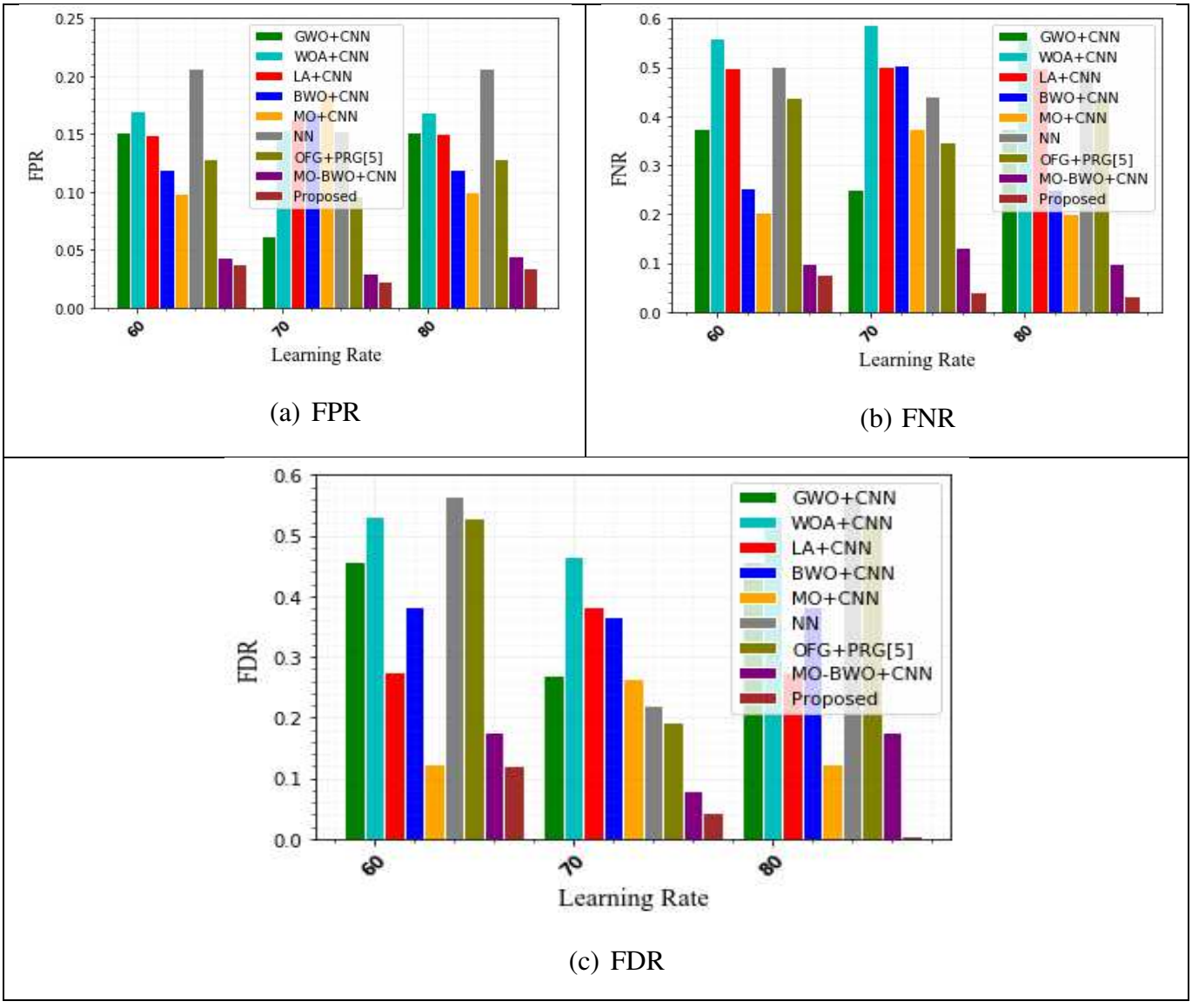
Figure (8) Simulation results of **(a)** Specificity **(b)** NPV and **(c)** MCC with varying learning rate.

Methodologies	**Learning rate 70 (%)**			**Learning rate 80 (%)**		
	Specificity (%)	**NPV (%)**	**MCC (%)**	**Specificity (%)**	**NPV (%)**	**MCC (%)**
GWO+CNN	88	92	72	82	82	58
WOA+CNN	82	80	50	80	80	39
LA+CNN	85	82	52	82	82	45
BWO+CNN	86	82	60	85	85	62
MO+CNN	87	84	58	90	90	75
NN	76	82	62	80	76	42
OFG+PRG [5]	80	81	73	86	82	57
MO-BWO+CNN	93	93	82	84	90	80
Proposed	**97**	**96**	**95.5**	**98**	**98.5**	**98.7**

Table 6. Comparison table for Specificity, NPV and MCC by varying learning rate 70 and 80.

Table 6 shows the Comparison table for Specificity, NPV and MCC by varying learning rate 70 and 80. Figure 9(a) is the FPR, FNR and FDR achieved values of proposed and the existing techniques. The minimum FPR value indicates the maximum efficiency in the proposed system here it achieves 0.03 as the FPR so the proposed method has higher efficiency than the existing approaches. If the learning rate is 80 the achieved FPR values of existing

GWO+CNN, WOA+CNN, LA+CNN, BWO+CNN, MO+CNN, NN, OFG+PRG [5], MO-BWO+CNN are: 0.15, 0.17, 0.15, 0.12, 0.10, 0.22, 0.13, 0.04 by comparing these existing values of FPR to the proposed the proposed GMM techniques achieves minimum FPR value so it is considered as much more efficient than others. Figure 9(b) shows the FNR values of existing and the proposed approaches comparison the lower value of FNR represents the higher efficiency in the video detection techniques. The FNR values for the existing GWO+CNN is 0.22, the WOA+CNN is 0.54, the LA+CNN is 0.50, the BWO+CNN is 0.50, the MO+CNN is 0.33, the NN is 0.42, the existing OFG+PRG [5] is 0.33, the MO-BWO+CNN is 0.12 and the FNR value for proposed GMM is 0.03 by considering the learning size 70. The FDR graphical notation for the proposed and the existing comparison is shown in the figure 9(c), here also higher efficiency is attains if the FDR rate is less. The proposed method reaches 0.04 as the FDR values and it is very low when compared with existing approaches. In the learning rate 70, the implemented GWO+CNN FDR rate is 0.23, the implemented WOA+CNN FDR value is 0.43, the FDR rate of existing LA+CNN attains 0.34, the existing BWO+CNN is 0.33, the MO+CNN FDR is 0.23, the NN method reaches 0.22, the OFG+PRG [5] FDR is 0.19, at last the previous MO-BWO+CNN technique gains 0.08 respectively.



(a) FPR

(b) FNR

(c) FDR

Figure (9): Performance measures of **(a)** FPR **(b)** FNR and **(c)** FDR with various learning rate.

Methodologies	Learning rate 70 (%)			Learning rate 80 (%)		
	FPR (%)	**FNR (%)**	**FDR (%)**	**FPR (%)**	**FNR (%)**	**FDR (%)**
GWO+CNN	0.06	0.23	0.28	0.15	0.38	0.48
WOA+CNN	0.15	0.59	0.48	0.17	0.59	0.55
LA+CNN	0.17	0.50	0.39	0.15	0.50	0.29
BWO+CNN	0.17	0.50	0.38	0.12	0.25	0.39
MO+CNN	0.18	0.33	0.28	0.10	0.20	0.13
NN	0.15	0.42	0.22	0.22	0.50	0.57
OFG+PRG [5]	0.09	0.32	0.19	0.13	0.42	0.55
MO-BWO+CNN	0.04	0.12	0.09	0.04	0.10	0.29
Proposed	**0.03**	**0.03**	**0.04**	**0.03**	**0.03**	**0.04**

Table 7: Comparison table for FPR, FNR and FDR by varying learning rate 70 and 80.

Figure (10): Forged output

Figure (11): Localized output

Table 7 shows the comparison table for FPR, FNR and FDR by varying learning rate 70 and 80. Figure 10 shows the Forged Output and Figure 11 Localized Output

5.3 Discussion

The research work processed out in this division is manipulated into four stages, namely key-frame extraction, pre-processing, multiple feature extraction and detection. The key-frame extraction play a dominant role in extracting key-frames from the videos. The next is pre-processing stage and here the RGB frames are converted into grayscale image. In this stage, some of the noises such as temporal noise, Gaussian noise is highly eliminated. Moreover, multi-features are extracted from frame video sequence to understand the nature of the video. Some of the features like SURF, PCA-HOG, MBFDF, CAF and PRG are extracted from the frames. Finally, the forgery happened in frames can detected and localized using ResNet152V2 with Bi-GRU based improved remora optimization model. The evaluation of the frame performance metrics are analyzed and compared with existing techniques such as GWO+CNN, WOA+CNN, LA+CNN, BWO+CNN, MO+CNN, NN, OFG+PRG and MO-BWO+CNN models. However, the existing methods attain low performance measures due to multiple disadvantages. The accurate detection and localization of video forgery are very difficult while using DL based CNN model. However, the extraction of spatial features and loss function reduction are efficient. The over fitting may occur while processing the implementation using CNN. So, the reliability of the CNN model reduced to 89%. So, in RNN network model highly affected due to noise in the video frames due to absence of pre-processing stage. Hence, this model degrades the accuracy to 92%.

The machine learning model uses some extraction approach to withdraw features based on spatiotemporal analysis but, this approach degrades the accuracy to 82% because of its complexity of training process. However, in KNN based machine learning technique uses SIFT approach to withdraw features from the frames. But, this method attains an accuracy of 86% due to low resolution outcome. In MMSP based traditional algorithms the forgery can be determined by correlation technique for each frames. But, this method degrades the accuracy of 76% due to noisy outcome. Finally, MSCL based video forgery detection technique utilizes some augmentation and segmentation approach to improve the quality of the output. However, this approach degrades the accuracy to 93% because of complex algorithm and time consuming process. When distinguished with aforementioned techniques, our proposed model yields better reliability, precision, sensitivity, specificity, FPR, FNR, F-measure, NPV, FDR and MCC consecutively. Our proposed approach uses improved remora optimization (IRO) algorithm to fine tune the hyper parameters and loss function reduction. We use the efficient GMM model for the purpose of key frame extraction and compare the performance with Otsu threshold and grey threshold approaches. This optimized deep learning model is inclined efficiently to detect and localize the object removal video forgery. The performance metrics shows the efficacy of our proposed model.

The evaluation metrics are inspected by changing the training percentage and the rate of training rate about 80%. By varying the learning rate, the performance can be analyzed and compared with the existing techniques. The existing optimization approach ROA are suffered due lack of robustness and may arise imbalance in exploration and exploitation strategies. But, the IROA strategy maintain good balance among exploration and exploitation ability. Moreover, it achieves flexible solutions with fine tune parameter and obtain efficient accuracy. In spite of its outstanding performance, the IROA is normalized with proposed network model for tuning parameters and loss function reduction.

6. Conclusion In digital video, object based modifications like integrating, eliminating or replacing the object are usually said to be as VF. In this paper, a DL based NN model is introduced for the detection and localization of video forgery. In recent times, detecting and

localizing the video forgery is a challenging task and need more advancements in the field of digital communication. In this research paper, the forged video can be processed under four stages namely, key-frame extraction, pre-processing, multi-feature extraction and detection. GMM model is used to extract frames from the video sequence. Pre-processing stage is emphasized to convert RGB frames into grayscale frames. In addition to this, ResNet152V2 with Bi-GRU is employed to detect and localize the forged video. Finally, improved remora optimization algorithm is manipulated to tune the hyper parameters and loss function reduction. Our proposed model reaches a better accuracy for the detection of VF. The performance analysis of the proposed model had been manifested via the performance metrics of reliability 96.17%, precision 96%, sensitivity 96.558%, F-measure 96.14%, MCC 0.92, specificity 96.58%, FPR 0.034, FNR 0.034, NPV 96% and FDR 0.04. In addition to this, the performance matrix for GMM model is compared with Otsu threshold and grey threshold. The outcome is based on MSE and PSNR and it attained about 27.95 and 104 respectively. In future work, various experiments will be conducted based on transfer learning approach for the better detection of forged video.

Abbreviations

VF: Video Forgery, DL: Deep Learning, VFD: video forgery detection, GMM: Gaussian mixture model, CMF: copy-move forgery, MMSP: multimedia signal processing, Bi-GRU: bi-directional gated recurrent unit, OA: optimization algorithm, SURF: speeded up robust features, PCA: principal component analysis, HOG: histogram of oriented gradient, MBFDF: model based fast digit feature, PRG: prediction residual gradient, OFG: optical flow gradient, ROA: remora optimization algorithm, AFM: Autonomous foraging mechanism.

Acknowledgements

Not applicable.

Authors' contributions

LK has found the proposed algorithms and obtained the datasets for the research and explored different methods discussed and contributed to the modification of study objectives and framework. Their rich experience was instrumental in improving our work. PKV has done the literature survey of the paper and contributed writing the paper. All authors contributed to the editing and proofreading. All authors read and approved the final manuscript.

Funding

Authors did not receive any funding for this study.

Availability of data and materials

In this work, for detection and localization of VF we take the sample videos from REWIND dataset and it is manipulated in the GUI interfaced. This dataset consist of total amount of 80 video sequence of about 320×240 pixel resolution and 30 fps frame-rate. The video sequence has been constructed initially from 10 original video sequence, that has been saved by low-end devices and narrow at the origin (by either MJPEG or H264 on the utilized device).

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no Competing interests.

Author details

1 Research Scholar, Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur-Dt, Andhra Pradesh.

2 Associate Professor, Department of Computer Science & Engineering, Koneru Lakshmaiah Education Foundation, Vaddeswaram, Guntur-Dt, Andhra Pradesh.

Reference:

[1] Bennett, Eric P., and Leonard McMillan. "Proscenium: a framework for spatio-temporal video editing." In *Proceedings of the eleventh ACM international conference on Multimedia*, pp. 177-184. 2003.

[2] Al-Sanjary, Omar Ismael, Ahmed Abdullah Ahmed, and Ghazali Sulong. "Development of a video tampering dataset for forensic investigation." *Forensic science international* 266 (2016): 565-572.

[3] Ali, Amir Hatem. "The power of social media in developing nations: New tools for closing the global digital divide and beyond." *Harv. Hum. Rts. J.* 24 (2011): 185.

[4] Bourouis, Sami, Roobaea Alroobaea, Abdullah M. Alharbi, Murad Andejany, and Saeed Rubaiee. "Recent advances in digital multimedia tampering detection for forensics analysis." *Symmetry* 12, no. 11 (2020): 1811.

[5] Yang, Quanxin, Dongjin Yu, Zhuxi Zhang, Ye Yao, and Linqiang Chen. "Spatiotemporal trident networks: detection and localization of object removal tampering in video passive forensics." *IEEE Transactions on Circuits and Systems for Video Technology* 31, no. 10 (2020): 4131-4144.

[6] Nabi, Syed Tufael, Munish Kumar, Paramjeet Singh, Naveen Aggarwal, and Krishan Kumar. "A comprehensive survey of image and video forgery techniques: variants, challenges, and future directions." *Multimedia Systems* (2022): 1-54.

[7] Soni, Badal, Pradip K. Das, and Dalton Meitei Thounaojam. "CMFD: a detailed review of block based and key feature based techniques in image copy-move forgery detection." *IET Image Processing* 12, no. 2 (2018): 167-178.

[8] Karnati, Mohan, Ayan Seal, Anis Yazidi, and Ondrej Krejcar. "LieNet: A Deep Convolution Neural Networks Framework for Detecting Deception." *IEEE Transactions on Cognitive and Developmental Systems* (2021).

[9] Bestagini, Paolo, Simone Milani, Marco Tagliasacchi, and Stefano Tubaro. "Local tampering detection in video sequences." In *2013 IEEE 15th international workshop on multimedia signal processing (MMSP)*, pp. 488-493. IEEE, 2013.

[10] Lin, Cheng-Shian, and Jyh-Jong Tsay. "A passive approach for effective detection and localization of region-level video forgery with spatio-temporal coherence analysis." *Digital Investigation* 11, no. 2 (2014): 120-140.

[11] Pandey, Ramesh Chand, Sanjay Kumar Singh, and K. K. Shukla. "Passive copy-move forgery detection in videos." In *2014 International conference on computer and communication technology (ICCCT)*, pp. 301-306. IEEE, 2014.

[12] Leo, Marco, G. Medioni, M. Trivedi, Takeo Kanade, and Giovanni Maria Farinella. "Computer vision for assistive technologies." *Computer Vision and Image Understanding* 154 (2017): 1-15.

[13] Hu, Fan, Gui-Song Xia, Jingwen Hu, and Liangpei Zhang. "Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery." *Remote Sensing* 7, no. 11 (2015): 14680-14707.

[14] Shelke, Nitin Arvind, and Singara Singh Kasana. "A comprehensive survey on passive techniques for digital video forgery detection." *Multimedia Tools and Applications* 80, no. 4 (2021): 6247-6310.

[15] Han, Junwei, Dingwen Zhang, Gong Cheng, Nian Liu, and Dong Xu. "Advanced deep-learning techniques for salient and category-specific object detection: a survey." *IEEE Signal Processing Magazine* 35, no. 1 (2018): 84-100.

[16] Yao, Ye, Yunqing Shi, Shaowei Weng, and Bo Guan. "Deep learning for detection of object-based forgery in advanced video." *Symmetry* 10, no. 1 (2018): 3.

[17] D'Avino, Dario, Davide Cozzolino, Giovanni Poggi, and Luisa Verdoliva. "Autoencoder with recurrent neural networks for video forgery detection." *Electronic Imaging* 2017, no. 7 (2017): 92-99.

[18] Sasikumar, R., K. Thaslima Nasreen, and C. Jeganathan. "Video Forgery Detection Using Deep Learning Techniques And Clustering Algorithms." *Studia Rosenthaliana (Journal for the Study of Research)* 12, no. 5 (2020).

[19] Kohli, Aditi, Abhinav Gupta, and Divya Singhal. "CNN based localisation of forged region in object-based forgery for HD videos." *IET Image Process.* 14, no. 5 (2020): 947-958.

[20] Fadl, Sondos, Qi Han, and Qiong Li. "CNN spatiotemporal features and fusion for surveillance video forgery detection." *Signal Processing: Image Communication* 90 (2021): 116066.

[21] Patel, Jatin, and Ravi Sheth. "An Optimized Convolution Neural Network Based Inter-Frame Forgery Detection Model-A Multi-Feature Extraction Framework."

[22] Soeleman, Moch Arief, Aris Nurhindarto, W. Karis, Muljono Farikh Al Zami, and R. Anggi Pramunendar. "Adaptive threshold for moving objects detection using gaussian mixture model." *Telkomnika* 18, no. 2 (2020): 1122-1129.