Journal of Ambient Intelligence and Humanized Computing (2021) 12:1197–1210
https://doi.org/10.1007/s12652-020-02164-y

ORIGINAL RESEARCH

Hybrid convolutional neural network (CNN) and long-short term memory (LSTM) based deep learning model for detecting shilling attack in the social-aware network

K. Vivekanandan[1] · **N. Praveena**[1]

Received: 3 November 2019 / Accepted: 30 May 2020 / Published online: 6 June 2020

Abstract
In social aware network (SAN) paradigm, the fundamental activities concentrate on exploring the behavior and attributes of the users. This investigation of user characteristic aids in the design of highly efficient and suitable protocols. In particular, the shilling attack introduces a high degree of vulnerability into the recommender systems. The shilling attackers use the reviews, user ratings and forged user generated content data for the computation of recommendation rankings. The detection of shilling attack in recommender systems is considered to be essential for sustaining their fairness and reliability. In specific, the collaborative filtering strategies utilized for detecting shilling attackers through efficient user behavior mining are considered as the predominant methodologies in the literature. In this paper, a hybrid convolutional neural network (CNN) and long-short term memory (LSTM)-based deep learning model (CNN–LSTM) is proposed for detecting shilling attack in recommender systems. This deep learning model utilizes the transformed network architecture for exploiting the deep-level attributes derived from user rated profiles. It overcomes the limitations of the existing shilling attack detection methods which mostly focuses on identifying spam users by designing features artificially in order to enhance their efficiency and robustness. It is also potent in elucidating deep-level features for efficiently detecting shilling attacks by accurately elaborating the user ratings. The experimental results confirmed the significance of the proposed CNN–LSTM approach by accurately detecting most of the obfuscated attacks compared to the state-of-art algorithms used for investigation.

Keywords Shilling attack · Deep learning model · Recommender systems · User profiles · User ratings · Convolutional neural network (CNN) · Long-short term memory (LSTM)

1 Introduction

The social aware network (SAN) has evolved as the modern paradigms due to the dramatic developments occurred in the area of wireless network and communication technologies over the recent decades (Zhang et al. 2013). SAN is one of the novel paradigms that exploits the social characteristics of network devices with increasing types and numbers of wireless mobile equipments (Gunes et al. 2014). SAN is potent in comprehensively exploring the social attributes of individuals such as human-to-human relationship, human-to-environment relationship, human-to-community and personal information (Zhang and Zhou 2012). The personal information in SAN paradigm includes the attributes, behavior and habits and many more related to an individual interacting on the network (Krizhevsky et al. 2017). However, the exploration and investigation enforced over the personal information of an individual aids in identifying their willingness and preferences (Zhang et al. 2017). The personal information and user ratings of recommender systems aids in highlighting the social behavior and sociability of human beings. However, the shilling attackers introduce vulnerability in SAN by exploiting these user data in a malicious manner (Wang et al. 2015).

Further, shilling attack is considered as a binary classification problem, since the user profile based on the classification result can be categorized into normal and attacker. Thus, a number of features are required for detecting attacks

✉ N. Praveena
praveenan@pec.edu

K. Vivekanandan
k.vivekanandan@pec.edu

[1] Department of Computer Science and Engineering, Pondicherry Engineering College, Puducherry, India

and some machine learning methods are essential for discriminating genuine users from attackers. The features such as item popularity, timestamps and ratings are the features that are extracted based on human engineering from a piece of user generated information. These features are considered to focus on some significant category of attacks that necessitate a high degree of knowledge costs. It is very complex for fully featuring the shilling profiles from a single piece of user-generated data that represents a single view due to the variability and diversity of strategies imposed by the attackers. Furthermore, the shilling attack detection process is considered as the imbalanced classification problem as the numbers of fake profiles are comparatively rare compared to the genuine ones. Hence, shilling attack detection schemes always offer inferior performance with different types of attacks that constitutes of low attack sizes.

In Burke et al. (2005), collaborative filtering-based detection of shilling attacks in SAN through the extraction of user behavior mining is considered as the important and frequently evaluated topic in the dimension of recommender systems. However, the issue of information overhead is constantly increasing with a corresponding rapid increase in the online information. The collaborative filtering-based detection schemes are successful in the process of filtering irrelevant information or predicting users' futuristic profile by judging the items based on the behavioral attitudes of users' neighbors, since they are potential in achieving remarkable success through item recommendation successfully into the computer applications (Deng et al. 2013). However, the collaborative filtering-based detection schemes are considered to be highly vulnerable to shilling attacks due to the characteristic properties of recommender systems. Shilling attacks refers to the process of simply injecting a diversified number of automatically generated profiles into the recommender system for increasing or decreasing the items rating score of the target (Mobasher et al. 2007). The previous research contributions propounded in the field of SAN confirmed that shilling attacks are responsible for minimizing the potentiality of the recommender systems (Ji et al. 2007). Therefore, the need for detecting shilling attacks have emerged as a momentous issue for facilitating effective and stabilization of recommender systems.

In general, the shilling attacks detection approaches propounded in the field of the SAN are categorized into three classes that include unsupervised, supervised and semi-supervised schemes (Patel et al. 2016). The unsupervised schemes contributed for detecting shilling attacks are broadly categorized into principal component analysis (PCA) and clustering schemes. The PCA-based detection approaches are determined to ensure better results in terms of identification. On the other hand, clustering-based schemes posses a simple and frank principle as they facilitate the process of clustering profiles through the enforcement of artificially designed features (Karthikeyan et al. 2016). However, the PCA-based detection approaches are prone to average over popular attack. Further, the clustering-based schemes are unstable since some of the legitimate users share common similarities to the spam users. Meanwhile, the typical examples of supervised approaches include support vector machine methods, knn-based classification methods and k-nearest neighbor methods. However, the supervised and un-supervised shilling attack detection schemes posses some shortcomings (Cao 2016). The classical supervised detection approaches are vulnerable to obfuscated attacks that include mixed attacks which integrated fake or spurious profiles obtained from multiple shilling attack strategies. Semi-supervised approaches are identified to be more stable compared to the aforementioned unsupervised and supervised detection approaches (Kapoor et al. 2018). However, they incur unbearable prolonged time in computation with increased complexity compared to the existing unsupervised and supervised detection methodologies. Furthermore, the majority of the existing techniques detect shilling attackers by utilizing artificially designed features. The artificially formulated features are least non-linear in characteristics with most of the features considered for attack detection is formulated for specific categories of attack models. As a whole, the classic manually designed features are not capable enough in handling the impacts of complicated or unknown attacks. Thus, shilling attack detection scheme with maximized stability, accessibility and efficiency are the urgent requirement in this context. The rapid development and attained success visualized and realized through the utilization of deep learning theory over the last decades in the area of speech recognition, image classification, hash tag recommendation and handwritten character classification have motivated its utilization in detecting shilling attacks. In addition, the research contribution by Zhang and Zhou (2015) for detecting shilling attacks through CNN also formed the secondary motivation behind the formulation of the proposed CNN–LSTM approach.

In specific, the deep neural network is considered to be potent enough in superior data modeling, when they handle a large dataset that pertain to attack profile of the shilling attackers that are determined from the recommendation systems. In specific, deep learning models constitute of a category termed feature extractor networks. They are responsible for classifying between genuine user profile and attacker profile. The primary objective of this deep learning model complete focuses on learning high level and deep characteristics that are most potentially useful for classification or target detection. Feature extractors can be implemented based on the computation of convolution between the data and specific filters that are followed by the operation of down-sampling for retaining only the most significant features into account. One of the well

known feature extractor networks is CNN that composed of alternating layers such as (1) convolution filters connected locally, (2) process of down sampling subsequently followed by fully connected layer (Softmax layer) that acts as a classifier. CNN is capable in providing the benefits of selecting good features from user rating profiles and LSTM is renowned for its potentiality in learning sequential data. CNN and LSTM is also determined to be potential shilling attack detectors as a standalone approach. Thus, an attempt is made for integrating the benefits of CNN and LSTM for extracting temporal and spatial data that can better improve the accuracy of detecting shilling attacks by training and testing the user and attack profiles in a predominant manner.

In the computer-vision domain, the most traditional and highly used ANN structure is the CNN, which pertains to the type of deep feed-forward neural network. This CNN with the local connection, convolution, pooling and weight sharing is determined to minimize the complexity of the network and the number of parameters considered for training purpose. CNN can be easily optimized and trained based on its inherent unique fault tolerance capability and strong robustness. However, CNN has the extreme limitation of identifying an attacker profile that changes over the time period based on the two-dimensional kernel. In particular, the channel output of CNN only possesses the potential features after the calculation of the filter. CNN is further estimated to be poor at detecting gradual changes in the user profile examined for attack profile determination. The shilling attack is a persistent attack whose features are recognizable only to a specific extent until they aggregate for some specific amount of time. Hence, LSTM network is integrated with CNN for identifying shilling attack. LSTM is the category of RNN with more computationally complex unit. Further, LSTM is capable of processing variable-length input and learns highly nontrivial long distance dependence easily, since it inherited forget gate. Thus, CNN–LSTM is suitably integrated for identifying the systematic change involved in the attack profile learning process.

The main contributions of the proposed CNN–LSTM approach are listed as follows:

(1) It is the first hybrid convolutional neural network (CNN) and long-short term memory (LSTM)-based deep learning model contributed for detecting shilling attack with maximized accuracy and effectiveness compared to the existing classical detection approaches.
(2) It is one of the predominant shilling attack detection schemes that utilizes automatically generated deep-level characteristic features for facilitating adaptive and robust environment for handling the majority of attacks that even includes an unknown attack strategy.

The remaining sections of this paper are structured as follows. Section 2 highlights the detailed literature review of some of the potential shilling attack detection approaches contributed in the recent years. Section 3 presents the details and architecture of the proposed CNN–LSTM model utilized for significant shilling attack detection. Section 4 demonstrates the efficiency and effectiveness of the proposed CNN–LSTM model determined based on experiments, evaluations and investigations. Section 5 summarizes the paper with major contributions and scope of future research.

2 Related work

In this section, the significant shilling attack detection techniques proposed over the recent years are presented with their merits and limitations.

An ensemble-based shilling attack detection scheme using multiple dimension concentrated automatic feature extraction process was propounded for achieving improved efficiency under different attack implementation strategies (Hao et al. 2019). This ensemble method at the first level, explored the user behaviors based on multiple dimensions that include item popularity, user ratings and user–user graph. Then, it utilized the merits of stacked auto encoder with denoise characteristics for automatically deriving user parameters under diversified compromised rates based on preprocessed data determined in multiple dimensions. Finally, the extracted features are potentially integrated based on PCA. The experimental results of this ensemble method with Amazon, NetFlix and Movielens datasets confirmed its significance in facilitating predominant detection of diversified shilling attacks. However, the automatic feature extraction in this ensemble approach is not capable enough to the maximum level of accuracy in the detection process. Then, an ordered item sequence-based ensemble method (OIS-EM) of detecting shilling attacks was contributed to determining the deviation between legitimate and attack user profiles Zhang and Chen (2015). This OIS-EM approach initially built genuine item sequences and ordered popular item sequences for constructing genuine and popular item rating sequences corresponding to each user profile. It utilized six important features for representing the properties of attack profiles. It divided the complete set of items into an ordered item sequence for integrating them with mutual information with the view to extract the remaining four features of attack profiles. Finally, it utilized the method of bootstrap re-sampling with a simple majority voting methodology is used for constructing primitive training sets to build a comprehensive ensemble framework. The experimental results of the OIS-EM approach confirmed its potentiality in enhancing precision with a sustained high recall value. However, the accuracy in detection facilitated by the OIS-EM

approach is comparatively low. A shilling attack detection scheme using rating time series and abnormal group user estimations was proposed for improving the detection accuracy (Zhou et al. 2018). This detection scheme utilized the merits of rating prediction strategy for investigating credibility evaluation in order to determine proximity-oriented predictions. It included different strategies to estimate the suspicious rating by quantifying target item investigation and suspicious time windows. It used the merits of examining data streams and suspicious rating time segments for the purpose of identifying users' genuineness. The experiments conducted through standard datasets confirmed the impact of this proposed model.

Trust features and time series analysis (TF-TSA)-based shilling attack detection mechanism was proposed for enhancing the precision by estimating the rating pattern deviation between the genuine and attack profiles (Xu and Zhang 2019). This TF-TSA-based detection scheme ignored the relationship existing between the users quantified in terms of trust. It utilized a reactive rating-based distribution and forecasting model based on items-oriented time series. It included the method of single exponential method for detecting suspicious items from the user profiles. It possessed the potential of segregating suspicious user profiles from genuine user profiles based on the integration of trust associations and rating patterns. It also derived the benefits of SVM to separate attack profiles from a collection of suspicious user profiles. The experiments conducted using Epinions and CiaoDVD datasets proved its predominance evaluated in terms of precision and recall. A shilling attack detection scheme using unsupervised learning was proposed using prior knowledge of user rating characteristics for achieving superior precision and recall under mitigation process (Cai and Zhang 2019). This unsupervised learning approach identified the target items and associated goals of the attacks by exploring the difference between rating tendencies related to each item. Then, it constructed a suspicious user set by investigating the user rating behavioral characteristics in the dimension of rating preference and interest preference. It estimated the user interest preference and its diversity based block entropy and simple entropy methodology. It also explored the memory associated with user preferences based on the strategy of self-correlation analysis. The experiments conducted using Amazon Review dataset, Netflix dataset and Movie Lens dataset confirmed its predominance in terms of classification accuracy, precision, recall and F-Score. However, the artificially utilized features used in this unsupervised learning approach are the main limitation, since it is considered to be least non-linear. A statistical method for shilling attack detection was proposed to explore the rating patterns of attack profiles Zhou et al. (2015). This statistic-based detection scheme used two factors, namely Mean Agreement to Rating Deviation (MARD) and Top Neighbors with Similarity Degree (TNS Deg) for attaining the objective of investigating rating patterns of attack profiles. The parameter of TNS Deg played an anchor role in detecting complicated attack models. In addition, MARD factor was the key factors for segregating attack profiles from genuine user profiles with maximum classification accuracy. The experiments of this statistic-based detection scheme were proved to be superior in F-Measure and classification accuracy.

Furthermore, Bayesian Classifier-based Collaborating Filtering Scheme (BC-CFS) was contributed for detecting shilling attacks in order to measure privacy and robustness against attack profiles (Batmaz and Polat 2017). This BC-CFS scheme extracted quality data for preventing the malicious entities from building fake profiles. It utilized six types of attack models during the detection of the shilling attack for identifying disguised user item profiles. The empirical results of BC-CFS confirmed its predominance, even under perturbed binary ratings, thus enhancing the precision and F-Measure. A Semi-Supervised Learning Method of Shilling Attack Detection (SEMI-SLM-SAD) was proposed for exploiting the merits of different kinds of data utilized by the attacks during the process of attack generation (Cao et al. 2012). This SEMI-SLM-SAD initially trained a small portion of labeled users based on Bayes classifier, which is then utilized by the EM-λ unlabeled users for classifying attack profiles from genuine user profiles. The experiments of SEMI-SLM-SAD conducted using MovieLens dataset proved its efficiency with respect to unsupervised and supervised learning-based detectors. The experiments results of SEMI-SLM-SAD also proved its superiority in detecting diversified shilling attacks on par with the BC-CFS and TF-TSA approaches.

In addition, An Integrated Perception Patterns and Social Network Search-based Shilling Attack Detection (IPP-SNS-SAD) was proposed for improving the rate of classification accuracy in genuine profiles and attack profiles in SAN (Zhu 2018). The perception patterns used in IPP-SNS-SAD completely focused on the merits of CNN and multi-agents. It is considered as the parallel integration approach of CNN and multi-agents for acquiring prior knowledge from the user profiles in order to identify the existence of attack profiles. The multi-agents of IPP-SNS-SAD was mainly responsible for integrating the properties of group and individual community agents towards shilling attack detection. The Perception Patterns were used in IPP-SNS-SAD for establishing refinements in the intention of trust with the cooperation of external factors that played a vital role in the categorization process. The experiments of this IPP-SNS-SAD scheme confirmed a superior enhancement in F-Measure and classification accuracy. Finally, CNN-based shilling attack detection scheme (CNN-SADS) was proposed with the merits of exploiting deep level features derived from user profiles

(Tong et al. 2018). This exploitation of deep level feature was achieved using transformed network structure. CNN-SADS approach preventing the issue of utilizing artificially designed features based on the incorporation of potential features associated with CNN. The experimental results of CNN-SADS confirmed its excellence in precisely detecting diversified number of obfuscated attacks compared to the contributed IPP-SNS-SAD, Semi-SLM-SAD and BC-CFS techniques for applying security in SAN.

A shilling attack detection technique using binary ratings-based collaborative filtering was proposed for classifying six well-known categories of attack models (Batmaz et al. 2019). It incorporated attributes that are specific to the model that is capable in handling fake profiles in the recommendation system. It inherited as classification-inspired methodology which is potent in extracting the maximum number of shill profiles that correlate with binary data prior to the process of recommendation. This detection scheme included four models-specific attributes and six generic attributes for detecting attack profiles in an effective manner. The empirical results of this binary rating-based collaborative filtering approach was determined to successfully detection profiles of attackers, even when the size of the attack and fillers are relatively low. Then, shilling attack detection scheme using slope one algorithm was proposed in recommendation system based on the fusion properties of user similarity and trusted data (Jiang et al. 2018). This slope one algorithmic approach incorporates three procedures. In the first procedure, trusted data is selected, which is then used for calculating the user similarity in the second procedure. In the third procedure, similarity with respect to the weight factor is included for enhancing the characteristics of the improved slope one algorithm in order to determine the resultant recommendation equation. The experiments of this slope one algorithm were conducted using Amazon data set and the results were identified to be more accurate than the conventional slope one algorithm. An integrated emotion and trust-based collaborative filtering-based recommendation strategy was proposed against the shilling attack (Guo et al. 2018). It included a methodology that inherited implicit and explicit satisfaction for mitigating the issue of sparsity. It also included the process of establishing trust associations between the users based on the principle of subjective and objective trust. It computed objective trust based on the similarity of opinion based on the combination of preference and rating similarity. It also estimated subjective trust based on the determination of familiarity visualized among the users through 6° of discrimination. It finally established the trust relationship based on the trusted number of neighbors determined from the objective user. It had a significant approach for excluding the malicious users from the list of neighbors by screening them based on the degree of emotional consistency determined between the users derived from implicit user behavior characteristics. The experimental results of this emotion and trust-based shilling attack detection scheme were confirmed to enhance the degree of recommendation accuracy with respect to data sparsity.

An item popularity and item correlation-based shilling attack detection scheme was proposed for differentiating malicious samples collected through the user profiles derived from a recommendation system (Chen et al. 2019). This shilling attack detection approach was identified to be capable of confirming real user profiles that maximizes the potentiality of disguise. It was propounded to explore the rating of item correlation with respect to real user profiles that are quite different from the samples that are generated from the existing shilling attack methodologies. The experimental analysis of this shilling attack detection scheme attained the highest ability in attack after the elimination of suspected user profiles that determined based on SVM and PCA strategies. Item distribution and rating behavior-based shilling attack detection scheme was proposed to minimize the risk which occurs during the extraction of diverse features from the user profiles (Yang and Cai 2016). It prevented the limitation of enhancing the detection performance that relies on limited features, since they cannot completely represent the genuine and attack profiles. It included a mapping model that determined the association between item distribution and rating behavior by exploiting the solution based on the usage of least squares. It also incorporated a trained model for effective detection of shilling attack based on the employment of the regression principle. The extensive experiments conducted for this detection scheme demonstrated its potential in detection accuracy.

3 Hybrid convolutional neural network (CNN) and long-short term memory (LSTM)-based deep learning model

The CNN is considered as the potent neural network that utilizes the operation of convolution rather than the classical matrix multiplication operation in more than one layer of the network. It is potentially incorporated for handling data that possess similar kinds of grid structures. These grid structures are more commonly used in attack detection, image processing and computer vision. Initially, the multidimensional user profile data are directly utilized as the low-level input of CNN. Then, the potential features of the user profiles are extracted in a layer-by-layer basis based on the operation of convolution and pooling. The significant features of the user profiles used for shilling attack detection possess the invariant properties of scaling, rotation and translation. The output layer pertaining to the classical CNN is fully connected with the hidden layer. This process of

feature fusion that considers the complete outputs of the convolution layer is considered to be very simple for the objective of the proposed model. However, the issues of CNN includes depends on the problems associated with unnecessary information determined by the kernels, multiple kernels determining similar information and usage of bad kernels. At this juncture, it is potent in extracting deeper level of user profile features and enhance the accuracy in recognition by improving the number of convolutional layers, convolutional kernels and pooling layers. However, it is obvious in undoubtedly leading to a complex network that increasing the computation cost with a high risk exposure of over fitting. In this context, LSTM being a time recurrent neural network is highly suitable in processing the problem of sequence with improved time dependence. In LSTM, the input feature tensor is forgotten selectively, input and output based on three structures of threshold are utilized. It is predominant in the process of filtering, empty input fusion, convolutional kernels-based extraction of unnecessary information and similar information such that the aforementioned significant feature information could be predominantly stored in the state cell for a prolonged amount of time. In this context, the algorithms that hybridize the benefits of CNN and LSTM was contributed in the literature for achieving predominant results in text analysis, voice recognition, gesture recognition, machine health condition prediction, attacker profile analysis, rainfall prediction, and other fields. The input of the LSTM is also considered as the batch of data profiles determined in time series. But, the problem of shilling attacker detection in user profiles faces the challenges in task identification. The user profiles of the benchmarked datasets are considered as the original input to the CNN–LSTM–DLM. It prevents the issue of sequence dependency. Thus, this proposed CNN–LSTM is contributed as the online monitoring process of the user profiles for detecting shilling attacks like random, bandwagon, average attack, AoP attack and Love/Hate attack by integrating the benefits of CNN and LSTM as mentioned in the architecture highlighted in Fig. 1.

This proposed CNN–LSTM is formulated for the recognition task of detecting shilling attacks by exploring the features of the user profiles. This proposed CNN–LSTM is capable enough in preventing the reliance of data reconstruction and feature extraction process by completely relying the merits of subjective consciousness and human expertise, since the feature extraction potential of CNN is self-learning and adaptive in its characteristics. It included multiple convolution kernels for the purpose of scanning the complete user profile for attaining redundant or spurious features of all the complete set of objects considered as candidates. In specific, the three-dimensional feature tensor output derived from the CNNs' last layer is elaborated into a single one-dimensional feature vector in the stage of feature hybridization. This single one-dimensional feature vector is completely extracted in the proposed CNN–LSTM through the inclusion of convolution kernels, which includes inessential information, likelihood information and some blank information, etc.,. Further, this feature vector is transformed into

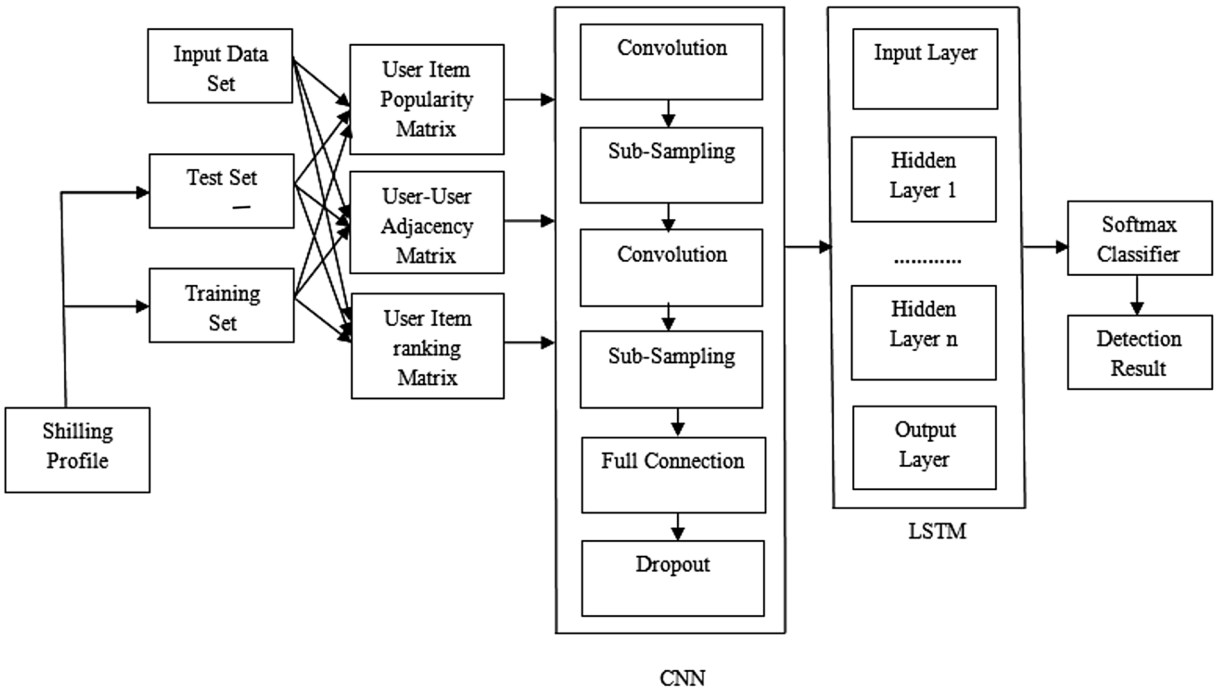

Fig. 1 The architecture of the proposed CNN–LSTM scheme

a two-dimensional space which is considered as the input to the LSTM. The individual row associated with feature matrix is considered as the primitive unit of hybridization. In the proposed CNN–LSTM, the feature information row is read at each time step and feature matrix is partitioned into diversified time steps to the recognition. Hence, the single user profile is converted into a sequential data by following the aforementioned strategy of implementation. The LSTM network plays a vital role in extracting the vital dependencies that lie between each individual row of feature matrix. This LSTM also plays an anchor role in filtering and integrating the features that are extracted from the CNN network. In addition, the excellence of the proposed CNN–LSTM network in detecting the shilling attack is depicted in Fig. 2.

In the time interval of the CNN–LSTM network designed for identifying shilling attackers over user profiles, the LSTM input at any time 't' comprises of unit state U_{t-1} and output O_{t-1} determined at time instant '$t-1$' with current time network input of x_t. Hence, it is clear that that the cell state and feature tensor could be hybridized and filtered through the utilization of three carefully formulated threshold structure. The design of the threshold structure forms the string base for extracting potential features from CNN, which could be stored in the cell state for a prolonged time with the additional merit of forgetting invalid features.

3.1 Parameter details and model implementation

From the Fig. 1, it is obvious that the proposed CNN–LSTM algorithm is mainly partitioned into two phases viz., (1) CNN-based feature extraction phase and (2) LSTM-based feature fusion phase.

3.2 CNN-based feature extraction phase

In the CNN-based feature extraction phase, the process of forward propagation of the user profile data is as follows. It is considered that the $k-1$ layer is an input layer or a pooling layer and k layer is the convolutional layer. The formula of computation associated with the k layer is presented in Eq. (1):

$$x_i^k = f\left(\sum_{j \in FM_i} x_i^{k-1} \times l_{ij}^k + b_j^k\right). \tag{1}$$

The term x_i^k on the left side of the aforementioned equation depicts the feature vector of the k layer. The right hand side of the equation highlights the summation and convolution operation for all related feature vector x_i^{k-1} of the $k-1$ layer and the ith convolutional kernel of the k layer. Further, an offset parameter is added and passed on the activation function $f(*)$. In this context, b is the offset, FM is the feature map pertaining to the upper layer with 'l' as the convolutional kernel.

Considering that the k layer is the down sampling or pooling layer and $k-1$ layer as the convolutional layer, the formula for the k layer is presented in Eq. (2):

$$x_i^k = f(\alpha_{DS_i^k} down(x_i^{k-1}) + b_j^k), \tag{2}$$

where, α_{DS} and $down(*)$ represent the down sampling coefficient and down sampling function respectively.

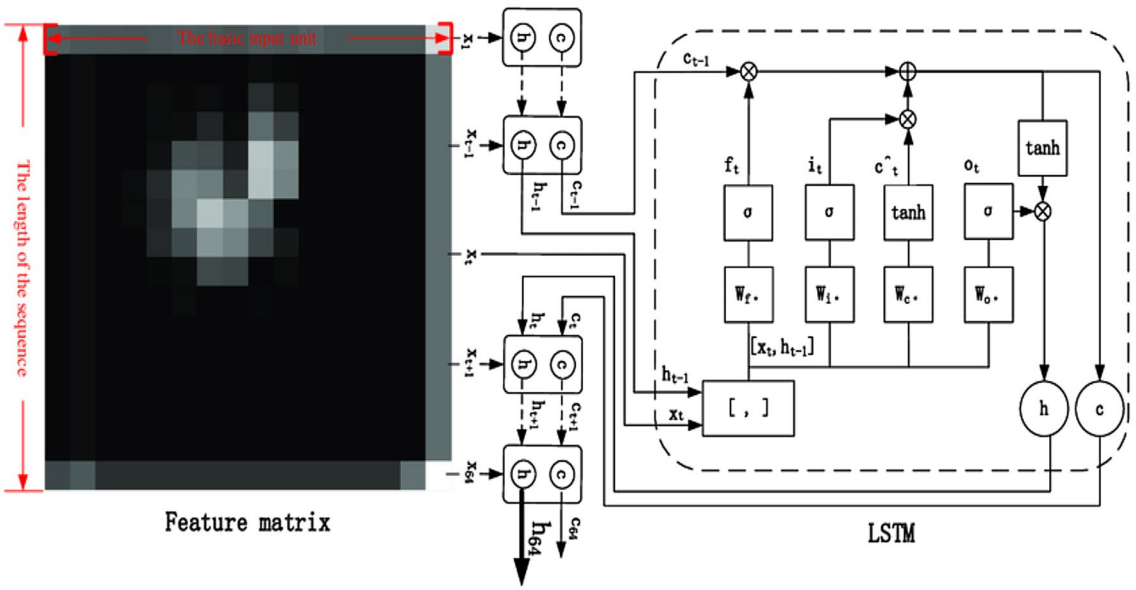

Fig. 2 Transformation of feature matrix in the proposed CNN–LSTM SCHEME

3.3 LSTM-based feature fusion phase

In this feature fusion phase, the network utilizes three threshold structures for controlling the state of the cells in order to preserve the importance of long-term memory. The significance of long short-term memory is completely based on two parameters Sm_t and Lm_t that presents short-term and long-term memory. The term $\sigma(*)$ used in Eqs. (3), (4) and (7) is the sigmoid function. At this juncture, the information is fully remembered when the Sigmoid function output is determined to be 1. The information is completely forgotten when the Sigmoid function output is determined to be 0. The portion of information need to be remembered when the Sigmoid function output is determined to lie between 0 and 1. Moreover the gate actually represents the fully connected layer of the proposed model with its input vector and its real vector output that lies between 0 and 1. In addition, LSTM utilizes the output vector $op_{(t)}$ of the gate by multiplying it with the vector which needs to be controlled. The forgetting gate f_t is responsible for estimating the amount of historical information which could be sustained in the long-term state. The input gate $ip_{(t)}$ is used for computing the amount of current network input information that could be possibly included into the long-term state. In addition, the output gate $op_{(t)}$ is used for controlling the amount of aggregated information available as the present output. The expressions of the aforementioned control parameters used in the LSTM-based feature fusion phase is presented as follows:

$$f_t = \sigma(W_{op} \times [h_{t-1}, x_t] + b_f) \tag{3}$$

$$ip_{(t)} = \sigma(W_{op} \times [h_{t-1}, x_t] + b_{in}) \tag{4}$$

$$Sm_t = \tanh(W_{in} \times [h_{t-1}, x_t] + b_{in}) \tag{5}$$

$$Lm_t = f_t \circ Lm_{t-1} + ip_{(t)} \circ Sm_t \tag{6}$$

$$op_{(t)} = \sigma(W_{op} \times [h_{t-1}, x_t] + b_{op}) \tag{7}$$

$$h_t = op_{(t)} \circ \tanh(Lm_t), \tag{8}$$

where, the symbol '$\circ$' represents the Hadamard product (element-wise product).

The aforementioned formulas pertain to the process of forward propagation towards the exploration of user profiles, which is considered as the input to the Softmax function. The kind of shilling attack is identified in terms of probability, once the processing output is derived from the Softmax function. In particular, the method of error back propagation is adapted by the deep neural network in the training stage of the proposed algorithm for updating the offsets and weights in an iterative manner until the maximum number of epochs is attained.

4 Experimental results and investigation

The predominance of the proposed CNN–LSTM model is tested and compared with the benchmarked CNN–SADS, IPP-SNS-SAD and SEMI-SLM-SAD schemes based on the publicly available Amazon review dataset, Movielens and Netflix dataset. The Amazon review dataset contains 1, 36, 785 items with 6,45,072 users and 12, 05,125 ratings, which is derived based on web crawling through Amazon.cn till August 20, 2012. However, the sample review dataset used comprised of 5050 users with 52,777 ratings derived over 17,610 items. The MovieLens 100 k dataset (http://grouplens.org/datasets/movielens/) comprises of 943 users and 1,00,000 ratings over 1682 movies. The Netflix dataset (http://www.netflixprize.com/) comprises of 4,80,000 users and randomly selected 1,00,000,000 ratings derived from anonymous users over 17,770 movies. The ratings in the Amazon review dataset, Movielens and Netflix dataset are integers that lie between 1 and 5, with 1 (dislike) and 5 (like) highlighting the lowest and highest ratings given by the users. The parameters considered for the implementation of the proposed CNN–LSTM model is presented in Table 1.

The performance of the proposed CNN–LSTM model and the benchmarked CNN–SADS, IPP-SNS-SAD and SEMI-SLM-SAD schemes are investigated based on precision, recall and F-measure based on Eqs. (9), (10) and (11). F-Measure is only used, since the length of the CNN and LSTM layers used for implementation is known in the prior (Tables 2, 3):

Table 1 Parameters used to implement CNN–LSTM model for shilling attack detection

Parameters	Considered values
Number of convolutional layers	1/2/3
Number of pooling layers	2
Fully connected layer	2
Number of neurons	250
Learning rate	0.6
Optimizer	Adam optimization
Activation function	ReLU
Training epochs	30
Size of the LSTM units	100
Number of filters	32
Dropout probability	0.5

Table 2 Features of the dataset utilized for investigating the proposed CNN–LSTM

Name of the dataset	Items	Users	Density	Rating
Amazon	2341	1104	7.84	120,000
MovieLens	1682	943	6.30	100,000
Netflix	2000	2000	1.09	40,000

$$Precision = \frac{True\_Positive}{(True\_Positive + False\_Positive)} \tag{9}$$

$$Recall = \frac{True\_Positive}{(True\_Positive + False\_Negative)} \tag{10}$$

$$F - Measure = \frac{2Precision \times Recall}{(Precision + Recall)} \tag{11}$$

Furthermore, the significance of the proposed CNN–LSTM is determined by comparing it with the state-of-art mechanisms such as CNN-SADS, IPP-SNS-SAD and SEMI-SLM-SAD schemes against different attacks which are individually elaborated over the three different datasets. Figure 3 demonstrates the performance of the proposed CNN–LSTM mechanism in terms of F-Measure over the compared CNN-SADS, IPP-SNS-SAD and SEMI-SLM-SAD schemes against attacks such as Random attack, Bandwagon attack, Average attack, AoP attack and Love/Hate attack with respect to Amazon review dataset. The F-Measure of the proposed CNN–LSTM mechanism against the attacks is determined to be greater than 99%, compared to the benchmarked CNN-SADS, IPP-SNS-SAD and SEMI-SLM-SAD schemes. This performance clearly portrays that the proposed CNN–LSTM mechanism is capable of detecting shilling attack profiles under most of the circumstances with bare false against any individual attack mechanism. The performance of the CNN-SADS is also determined to be significant against all the five kinds of obfuscated attack as its F-Measure is also greater than 0.9, but comparatively less than the proposed CNN–LSTM mechanism. On the other hand, IPP-SNS-SAD and SEMI-SLM-SAD schemes performed not so well as the F-Measure struggled to reach 35% and 23% which is considered to exhibit an unstable outcome.

Figure 4 demonstrates the performance of the proposed CNN–LSTM mechanism in terms of F-Measure over the compared CNN-SADS, IPP-SNS-SAD and SEMI-SLM-SAD schemes against attacks such as Random attack, Bandwagon attack, Average attack, up the attack and Love/Hate attack with respect to MovieLens data set. The F-Measure of the proposed CNN–LSTM mechanism against the attacks with MovieLens data set is determined to be greater than 99%, compared to the benchmarked CNN-SADS, IPP-SNS-SAD and SEMI-SLM-SAD schemes. This performance clearly portrays that the proposed CNN–LSTM mechanism is capable of detecting shilling attack profiles under most of the circumstances with bare false against any individual attack mechanism. The performance of the CNN-SADS is also determined to be significant against all the five kinds of obfuscated attack as its F-Measure is also greater than 0.9, but comparatively less than the proposed CNN–LSTM mechanism. On the other hand, IPP-SNS-SAD and SEMI-SLM-SAD schemes performed not so well as the F-Measure struggled to reach 41% and 29% which is considered to exhibit an unstable outcome.

Figure 5 demonstrates the performance of the proposed CNN–LSTM mechanism in terms of F-Measure over the compared CNN-SADS, IPP-SNS-SAD and SEMI-SLM-SAD schemes against attacks such as Random attack, Bandwagon attack, Average attack, up the attack and Love/Hate attack with respect to NetFlix data set. The F-Measure of the proposed CNN–LSTM mechanism against the attacks with NetFlix data set is determined to be greater than 99%, compared to the benchmarked CNN-SADS, IPP-SNS-SAD and SEMI-SLM-SAD schemes. This performance clearly portrays that the proposed CNN–LSTM mechanism is capable of detecting shilling attack profiles under most of the circumstances with bare false against any individual attack mechanism. The performance of the CNN-SADS is also determined to be significant against all the five kinds of obfuscated attack as its F-Measure is also greater than 0.9, but comparatively less than the proposed CNN–LSTM

Table 3 Different attack models considered during the implementation of the proposed CNN–LSTM

Type of attack	Filler item ratings	Rating	Selected items	Rating	Target items	I_{φ}
Segment	3	2	1	1	2	1
Bandwagon	2	2	1	1	2	1
Average	3	2	1	1	2	1
Random	2	3	2	2	2	2
Mixed	1	2	1	2	1	1
AoP	1	1	1	3	1	1

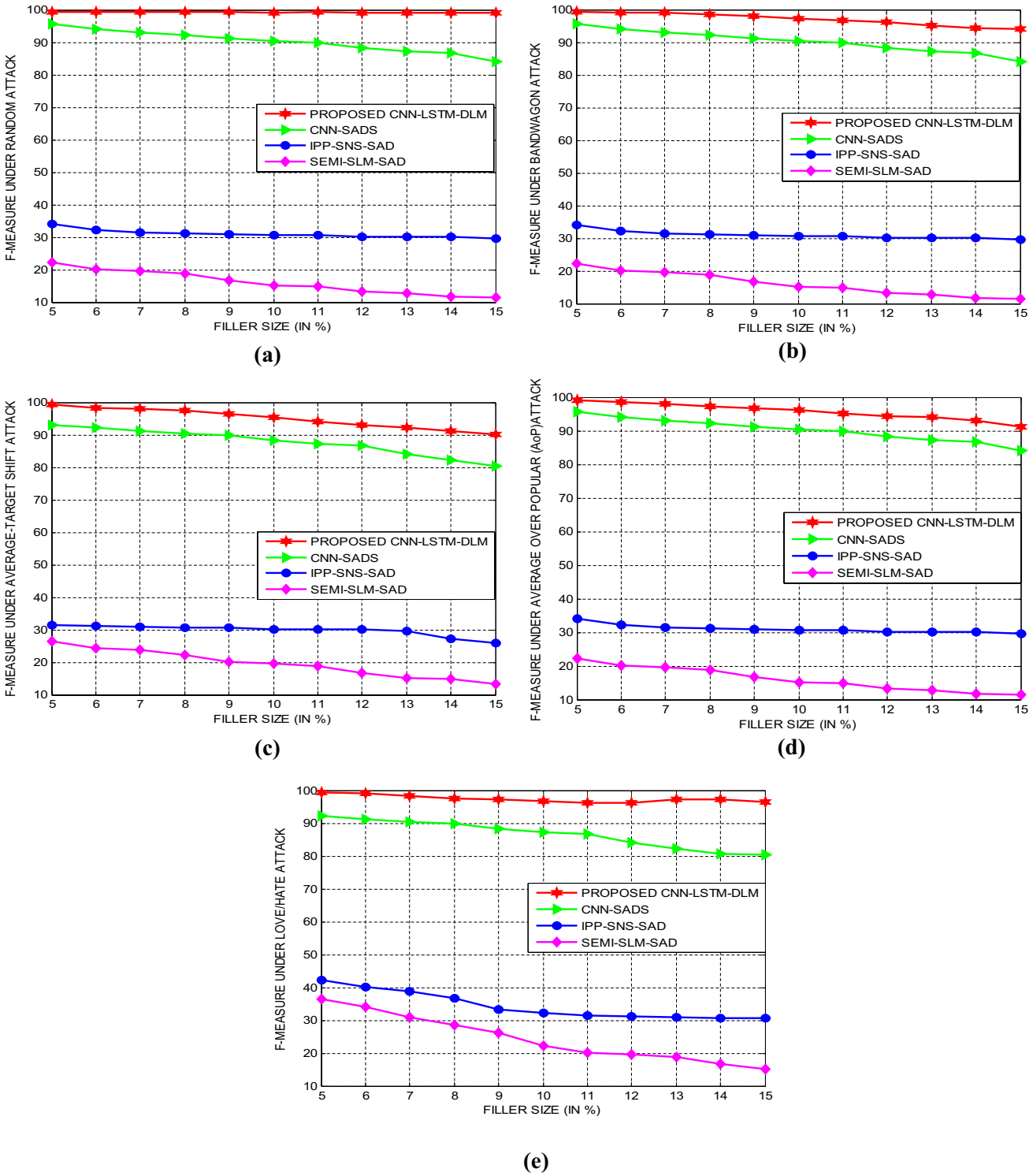

Fig. 3 Comparison of detectors using F-Measure with Random attack, Bandwagon attack, Average attack, AoP attack and Love/Hate attack with Amazon dataset

mechanism. On the other hand, IPP-SNS-SAD and SEMI-SLM-SAD schemes performed not so well as the F-Measure struggled to reach 39% and 28% which is considered to exhibit an unstable outcome.

In addition, the superior performance of the proposed CNN–LSTM mechanism and the benchmarked CNN-SADS, IPP-SNS-SAD and SEMI-SLM-SAD schemes are evaluated using classification accuracy, classification time and shilling

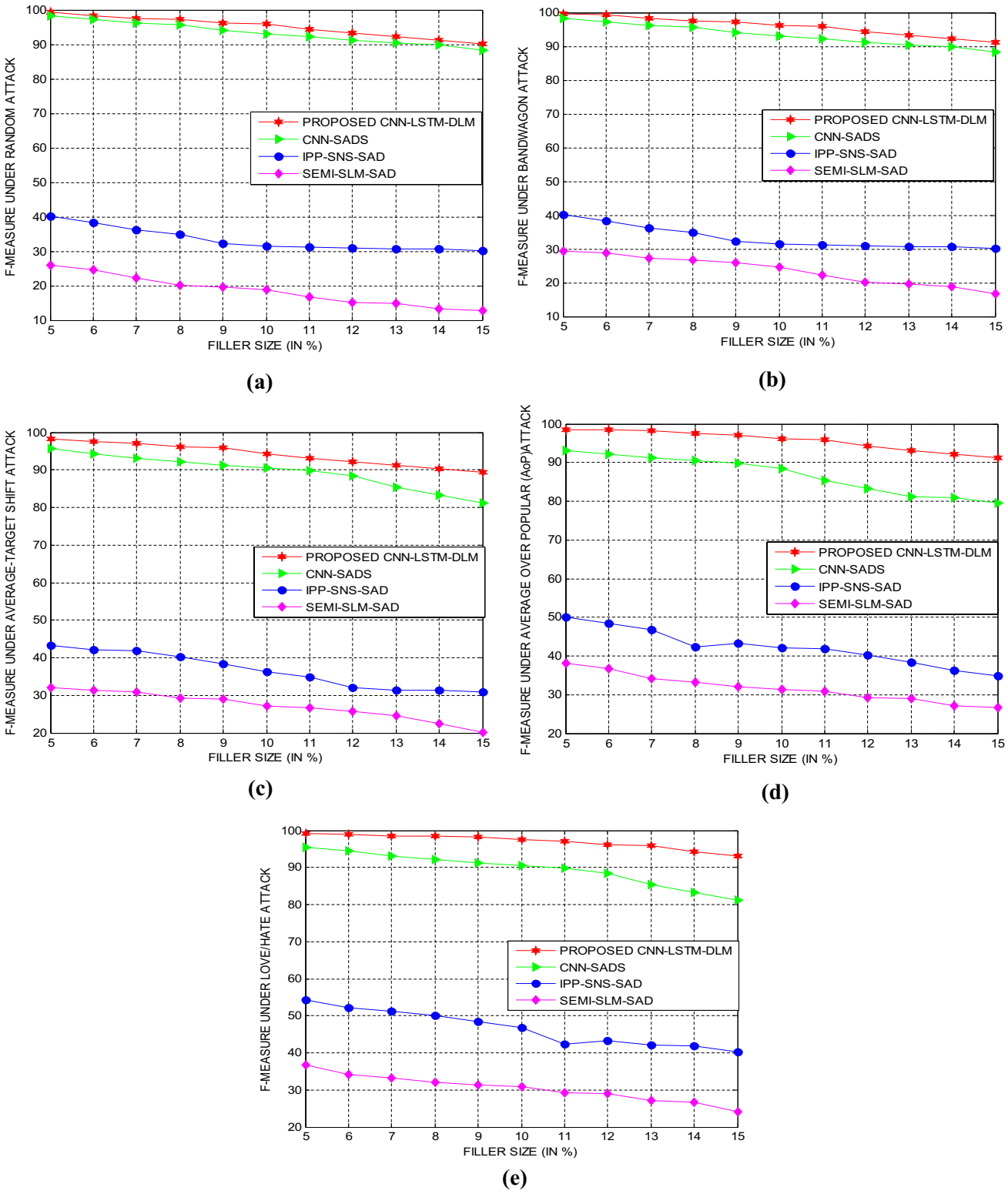

Fig. 4 Comparison of detectors using F-Measure with Random attack, Bandwagon attack, Average attack, AoP attack and Love/Hate attack with MovieLens dataset

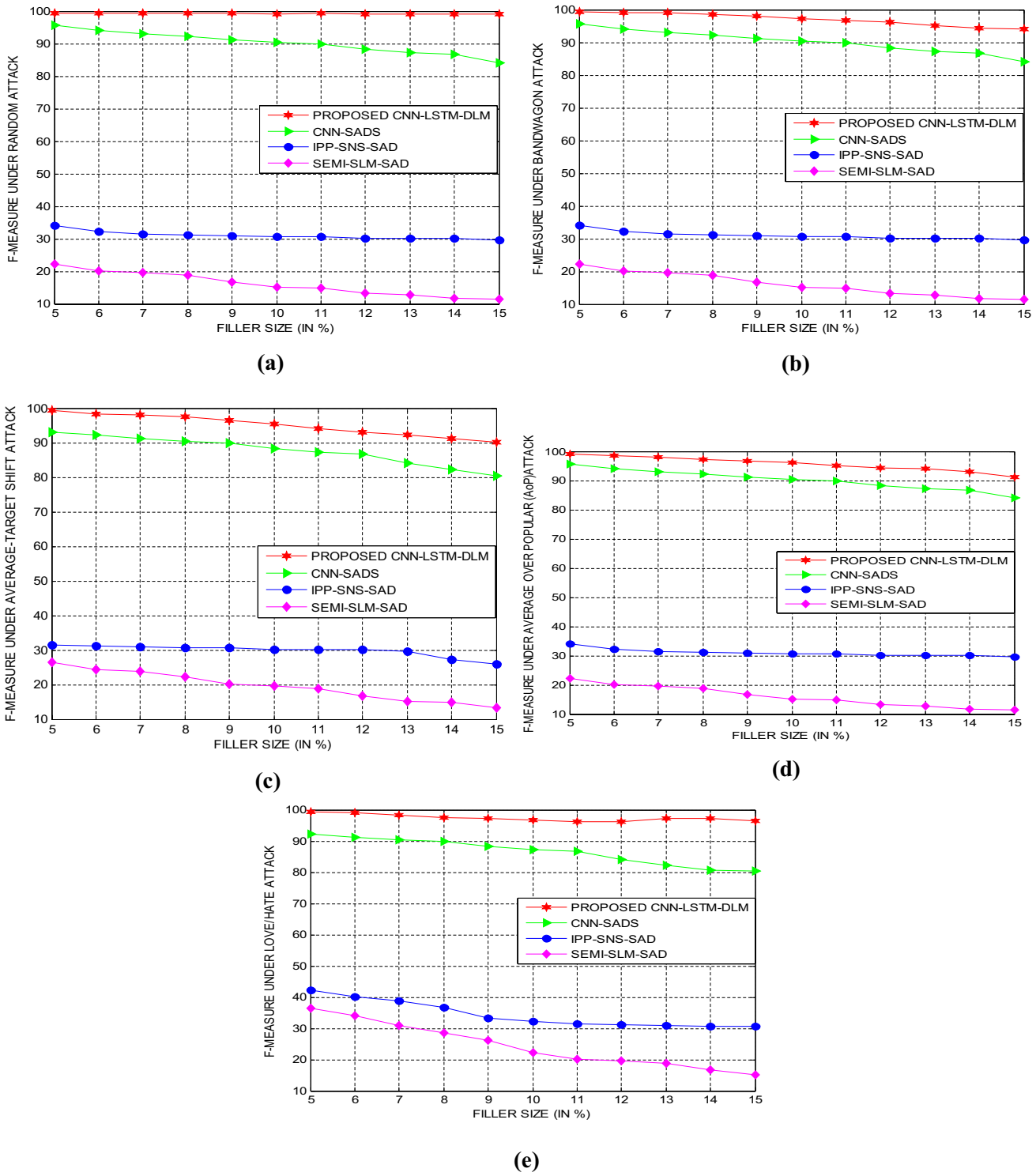

Fig. 5 Comparison of detectors using F-Measure with **a** Random attack, **b** Bandwagon attack, **c** Average attack, (d) AoP attack and **e** Love/Hate attack with Netflix dataset

attack detection rate are presented in Tables 4, 5 and 6 as follows.

The classification accuracy of the proposed CNN–LSTM mechanism is determined to be enhanced by an average margin of 4.21%, 5.32% and 6.74%, superior to the benchmarked CNN-SADS, IPP-SNS-SAD and SEMI-SLM-SAD schemes. The classification time of the proposed CNN–LSTM mechanism is determined to be enhanced by

Table 4 Proposed CNN–LSTM mechanism-classification accuracy

Mechanism	Percentage in classification accuracy (under different profiles)				
	20	40	60	80	100
CNN–LSTM	94.29	93.21	93.11	92.43	91.21
CNN-SADS	86.54	84.21	82.84	81.11	80.03
IPP-SNS-SAD	82.12	81.32	80.62	78.92	76.54
SEMI-SLM-SAD	80.74	78.54	75.41	73.32	71.29

Table 5 Proposed CNN–LSTM mechanism-classification time

Mechanism	Classification time (under different profiles)				
	20	40	60	80	100
CNN–LSTM	0.0000925	0.0001345	0.0001432	0.0001521	0.0001672
CNN-SADS	0.0001123	0.0001673	0.0001783	0.0001812	0.0001894
IPP-SNS-SAD	0.0001189	0.0001712	0.0001892	0.0001912	0.0001982
SEMI-SLM-SAD	0.0001213	0.0001768	0.0001898	0.0001921	0.0001988

Table 6 Proposed CNN–LSTM mechanism-shilling attack detection rate

Mechanism	Shilling attack detection rate (under different profiles)				
	20	40	60	80	100
CNN–LSTM	98.12	97.28	96.82	96.12	95.42
CNN-SADS	95.21	94.21	93.32	92.12	90.54
IPP-SNS-SAD	93.88	92.12	91.21	90.56	88.42
SEMI-SLM-SAD	92.21	90.46	89.64	88.42	85.92

an average margin of 6.82%, 7.94% and 8.12%, superior to the benchmarked CNN-SADS, IPP-SNS-SAD and SEMI-SLM-SAD schemes. In addition, the shilling attack detection rate of the proposed CNN–LSTM mechanism is determined to be enhanced by an average margin of 4.58%, 6.92% and 8.21%, superior to the benchmarked CNN-SADS, IPP-SNS-SAD and SEMI-SLM-SAD schemes.

5 Conclusions

In this paper, CNN–LSTM is contributed as an attempt for detecting shilling attack in recommendation systems as it is an imperative and impressive problem in the research field of SAN paradigm. This proposed CNN–LSTM incorporated the benefits of the transformed network architecture for utilizing the properties of deep-level features extracted from user rated profiles. It utilized the hybridization of CNN and LTSM-based neural networks for automatic extraction of discriminative and representative deep-level features. Thus, it prevented the shortcomings of the current shilling attack detection approaches that maximally concentrated on the process of confirming spam users through the inclusion of manually formulated features. The utilized transformed network architecture is constructed with single convolution, single transformation, single pooling and single output layer. Further, the conversion of user rating profiles into matrices is achieved through matrix transformation layer which is the essential process to satisfy the major requirement of CNN. Furthermore, the pooling layer is utilized in the proposed scheme for reducing the representation and attaining secondary level feature extraction. The comparative investigation of the proposed CNN–LSTM with the benchmarked state-of-art techniques evaluated with six different attack methodologies proved its predominance under major circumstances. The proposed CNN–LSTM scheme is confirmed to facilitate an accuracy rate of more than 98% independent of the kind of attack strategies considered during shilling attack detection. In the near future, it is also planned to devise a Hybrid Convolutional Neural Network and Restricted Boltzmann Machine (RBM) based deep learning model for training the deep neural network with huge and highly diversifies datasets through the inclusion of potent training strategies.

References

Batmaz Z, Polat H (2017) Designing shilling attacks on disguised binary data. Int J Data Min Model Manag 9(3):185

Batmaz Z, Yilmazel B, Kaleli C (2019) Shilling attack detection in binary data: a classification approach. J Ambient Intell Hum Comput 2(1):23–32

Burke R, Mobasher B, Bhaumik R, Williams C (2005) Segment-based injection attacks against collaborative filtering recommender systems. In: Fifth IEEE international conference on data mining, vol 1, (ICDM'05), pp 45–53

Cai H, Zhang F (2019) Detecting shilling attacks in recommender systems based on analysis of user rating behavior. Knowl Based Syst 177(3):22–43

Cao L (2016) Non-iid recommender systems: a review and framework of recommendation paradigm shifting. Engineering 2(2):212–224

Cao J, Wu Z, Mao B, Zhang Y (2012) Shilling attack detection utilizing semi-supervised learning method for collaborative recommender system. World Wide Web 16(5–6):729–748

Chen K, Chan PPK, Zhang F, Li Q (2019) Shilling attack based on item popularity and rated item correlation against collaborative filtering. Int J Mach Learn Cybern 10(7):1833–1845

Deng P, Zhong J et al (2013) Recommendation-based anti-attack trust model on E-commerce. J Comput Appl 33(12):3490–3493

Gunes I, Kaleli C, Bilge A, Polat H et al (2014) Shilling attacks against recommender systems: a comprehensive survey. Artif Intell Rev 42(4):767–799

Guo L, Liang J, Zhu Y, Luo Y, Sun L, Zheng X (2018) Collaborative filtering recommendation based on trust and emotion. J Intell Inf Syst 53(1):113–135

Hao Y, Zhang F, Wang J, Zhao Q, Cao J (2019) Detecting shilling attacks with automatic features from multiple views. Secur Commun Netw 2019(2):1–13

Ji A, Yeon C, Kim H, Jo G (2007) Distributed collaborative filtering for robust recommendations against shilling attacks. Adv Artif Intell 2(1):14–25

Jiang L, Cheng Y, Yang L, Li J, Yan H, Wang X (2018) A trust-based collaborative filtering algorithm for e-Commerce recommendation system. J Ambient Intell Hum Comput 10(8):3023–3034

Kapoor S, Gupta V, Kumar R (2018) An obfuscated attack detection approach for collaborative recommender systems. J Comput Inf Technol 26(1):45–56

Karthikeyan P, Selvi ST, Neeraja G, Deepika R, Vincent A, Abinaya V (2016) Prevention of shilling attack in recommender systems using discrete wavelet transform and support vector machine. In: Eighth international conference on advanced computing, vol 1, (ICoAC), pp 34–43

Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. Commun ACM 60(6):84–90

Mobasher B, Burke R, Bhaumik R, Sandvig JJ (2007) Attacks and remedies in collaborative recommendation. IEEE Intell Syst 22(3):56–63

Patel K, Thakkar A, Shah C, Makvana K (2016) A state of art survey on shilling attack in collaborative filtering based recommendation system. In: Proceedings of first international conference on information and communication technology for intelligent systems, vol 1, pp 377–385

Tong C, Yin X, Li J, Zhu T, Lv R, Sun L, Rodrigues JJPC (2018) A shilling attack detector based on convolutional neural network for collaborative recommender system in social aware network. Comput J 61(7):949–958

Wang W, Zhang G, Lu J (2015) Collaborative filtering with entropy-driven user similarity in recommender systems. Int J Intell Syst 30(8):854–870

Xu Y, Zhang F (2019) Detecting shilling attacks in social recommender systems based on time series analysis and trust features. Knowl Based Syst 178(4):25–47

Yang Z, Cai Z (2016) Detecting abnormal profiles in collaborative filtering recommender systems. J Intell Inf Syst 48(3):499–518

Zhang F, Chen H (2015) An ensemble method for detecting shilling attacks based on ordered item sequences. Secur Commun Netw 9(7):680–696

Zhang F, Zhou Q (2012) A meta-learning-based approach for detecting profile injection attacks in collaborative recommender systems. J Comput 7(1):67–74

Zhang F, Zhou Q (2015) Ensemble detection model for profile injection attacks in collaborative recommender systems based on BP neural network. IET Inf Secur 9(1):24–31

Zhang Z, Kulkarni SR et al (2013) Graph-based detection of shilling attacks in recommender systems. In: IEEE International workshop on machine learning for signal processing, vol 2, pp 12–18

Zhang Q, Wang J, Huang H, Huang X, Gong Y (2017) Hashtag recommendation for multimodal microblog using co-attention network. In: Proceedings of the twenty-sixth international joint conference on artificial intelligence, vol 2, pp 56–63

Zhou W, Wen J, Koh YS, Xiong Q, Gao M, Dobbie G, Alam S (2015) Shilling attacks detection in recommender systems based on target item analysis. PLoS One 10(7):e0130968–e0130968

Zhou W, Wen J, Qu Q, Zeng J, Cheng T (2018) Shilling attack detection for recommender systems based on credibility of group users and rating time series. PLoS One 13(5):e0196533–e0196533

Zhu L (2018) A novel social network measurement and perception pattern based on a multi-agent and convolutional neural network. Comput Electr Eng 66(2):229–245