IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 1, 2021, accepted August 10, 2021, date of publication August 16, 2021, date of current version August 20, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3105000

DarkDetect: Darknet Traffic Detection and Categorization Using Modified Convolution-Long Short-Term Memory

MUHAMMAD BILAL SARWAR, MUHAMMAD KASHIF HANIF, RAMZAN TALIB, MUHAMMAD YOUNAS, AND MUHAMMAD UMER SARWAR

Department of Computer Science, Government College University, Faisalabad 38000, Pakistan

Corresponding author: Muhammad Kashif Hanif (mkashifhanif@gcuf.edu.pk)

ABSTRACT Darknet is commonly known as the epicenter of illegal online activities. An analysis of darknet traffic is essential to monitor real-time applications and activities running over the Darknet. Recognizing network traffic bound to unused Internet addresses has become undeniably significant for identifying and examining malicious activities on the internet. Since there are no authentic hosts or devices in an unused address block, any observed network traffic must be the aftereffect of misconfiguration from spoofed source addressed and other frameworks that monitor unused address space. However, the recent advancements in artificial intelligence allow digital systems to detect and identify darknet traffic autonomously. In this paper, we propose a generalized approach for darknet traffic detection and categorization using Deep Learning. We examine the state-of-the-art complex dataset, which provides excessive information about the darknet traffic and perform data preprocessing. Next, we analyze diverse feature selection techniques to select optimal features for darknet traffic detection and categorization. We apply fine-tuned machine learning (ML) algorithms which include Decision Tree (DT), Gradient Boosting (GB), Random Forest Regressor (RFR), and Extreme Gradient Boosting (XGB) on selected features and compare the performance. Next, we apply modified Convolution-Long Short-Term Memory (CNN-LSTM) and Convolution-Gradient Recurrent Unit (CNN-GRU) deep learning techniques to recognize the network traffic more accurately. The results demonstrate that the proposed approach outperforms the existing approaches by yielding the maximum accuracy of 96% of darknet traffic detection and 89% of darknet traffic categorization through XGB as a feature selection approach and CNN-LSTM a recognition model.

INDEX TERMS Characterization, darknet traffic, deep learning, encrypted traffic, machine learning, Tor, VPN.

I. INTRODUCTION

Network traffic monitoring and analysis are challenging since it helps to improve network performance, minimize your attack surface, enhance security, and improve the resources management [1]–[3]. Network traffic trade in crimes such as viruses, contract killers, poisons, and drugs happens through the Darknet [4]. The address space which is not used on the internet is called darknets or network telescopes. Darknet traffic is not speculated over the internet to interact with other computers and only passively accepts incoming packets without generating outgoing packets [5]. Tor is a virtual computer network, allows users to gain access to hidden Darknet resources. This technique implements second-generation onion routing. Basically, in this routing, a circuit is built incrementally, one hop by one hop. Usually, the darknet traffic is treated as misconfiguration because the host is acting as illegitimate. A deep analysis of darknet traffic is significantly essential to monitor real-time applications. The analysis of darknet traffic provides complete information to the cybersecurity specialists, and other IT operators about the services exploited by attackers or vulnerable to some attack [6]–[9]. Researchers now focus on analyzing the darknet traffic, specifically detecting Tor applications to determine the malicious activities [10]–[14]. To achieve the detection objective, authors used ML, and Deep Learning (DL)

The associate editor coordinating the review of this manuscript and approving it for publication was Diana Gratiela Berbecaru.

techniques [15]–[18]. Authors in [5], [19], [20] explored whether graph mining techniques can help to uncover such macroscopic coordinated events in darknet traffic. The darknet traffic is getting complex daily, and malicious activities such as physical threats, sales data theft, fraudulent activity, phishing attacks, and scams, DDoS attacks, illicit links, Illicit Drugs, and terrorism vary day by day [21]–[24].

Limitations of Existing Study: The state-of-the-artwork has been done by [15] on the detection and categorization of darknet traffic and taken as the base paper. The dataset used in this work contains a large number of missing values that cannot be ignored. The authors in this study [15] have not explained how they handled this limitation. There exist multiple recommended methods to handle missing values to produce meaningful information. The other limitation in this work is that the authors selected 22 features from 61 extracted features using only one feature selection technique. The deep learning approach tested on 15 best features out of 22 selected features that seem biased. Furthermore, the overall f-score is 86%, but the f-score of traffic categories Browsing, E-mail, File-Transfer, and VOIP are 51%, 67%, 75%, and 61% respectively, which are significantly less, that affects the overall performance of the recognition model. To the best of our knowledge, this is the only other baseline work in the current literature. Also, we used the dataset of that study through a fair comparison with them to make sense. Other studies on Darknet traffic used limited or outdated datasets.

This paper answers the following research questions. **RQ1:** The interpretation of models detecting network traffic heavily depend upon all network feature. How is a feature selection technique affect the performance of the detection model? **RQ2:** How to assess the impacts of data volume and dataset balance in using a deep learning approach to detect darknet traffic? **RQ3:** How to detect darknet traffic accurately at multi-level classification as Tor, Non-Tor, VPN, Non-VPN and traffic categories: Audio-Stream, chat, mail, p2p, VOIP, Video-Stream, Transfer, and Browsing using custom-tuned machine learning and deep learning predictive models?

Keeping in sight the above research questions, this paper makes the following contributions:

1) Propose a customized LSTM based CNN (CNN-LSTM) deep learning approach for detecting and identifying darknet network traffic.
2) Apply pre-processing data operations, including data imputations, feature selection, and data balancing.
3) Provide a set of customized parameters for accurately recognizing traffic type and application categories using CNN-LSTM with promising Accuracy, Precision, Recall, and F-score.
4) Present comparative results using various machine learning, deep learning techniques, and state-of-the-art to show the effectiveness of the proposed approach.

The paper's remainder is structured as follows: Section II demonstrates the relevant work to darknet traffic analysis and detection. Section IV provides the proposed approach for darknet traffic analysis and detection. Section V provides the evaluation and results. Section VI summarizes the paper and provides the future direction.

II. RELATED WORK

This section presents the detailed related work based on darknet traffic and activities monitoring and exploration. The related work methods are composed of the latest artificial intelligence techniques and other safety networking frameworks.

The authors in [25] proposed a system that tracks port scanning action patterns between several explored ports and classify basic configurations of the examined cluster of ports. This approach is claimed to be completely automatic with the latest techniques of graph modeling and text mining. It provides complete information to the cybersecurity specialists and different information technology (IT) operators about the services exploited by attackers or vulnerable to some attack. This technique stands out to be more useful since it becomes easy to identify the targets of attacks. The proposed system is tested on the massive dataset of Darknet or an internet telescope.

The researchers in [16] highlighted the difficulties that security agencies face to track the criminal activities going on the Darknet. Analyzing the images in a massive amount on the Darknet is a time-consuming task and still not an efficient method. To encounter this issue, the researchers study different automated image classification methods which are based on Semantic Attention Keypoint Filtering (SAKF) and proposed an approach that can omit the non-relevant topographies at the deep pixel-level such that all the pixels are being filtered out which are not relevant to the forensic process. This system is achieved by compounding the salience maps with Bag of Visual Words (BoVW). The researchers tested out their system on a custom-made dataset that contains the Tor image. Their finding concludes that MobileNet v1, Resnet50, and BoVW, composed of dense SIFT descriptors, stand out to be more proficient than other approaches. The authors in [10] proposed a machine learning-based threat identification system in which the machine learning classifiers are trained on darknet traffic. The network traffic flow over the Darknet is either considered to be malicious or wrongly configured. The darknet traffic flow comprises multiple intimidations, including DDoS attacks, botnets, spoofing, probes, and scanning attacks. The researchers examined the darknet network traffic flow configured at SURFnet and mined other network properties of the traffic flow. The researchers applied a supervised machine learning algorithm Random Forest and a concept drift detector that shows the system is efficiently capable of detecting benign and malign traffic and proficiently detecting hidden threats over the network.

The researchers in [17] emphasized the difficulties faced by the security investigators and law enforcers while monitoring and examining the network traffic flow at the Darknet. A renowned method for semantical analysis of market listing is top modeling, but this technique is shows lag in the capability to record the visual images found in those listing. The

TABLE 1. Comparative summary of state-of-the-art methods for Darknet traffic recognition.

References		[15]	[25]	[16]	[10]	[17]	[18]	[26]	[27]	[12]	[11]	[28]
Approach	Port Scanning	-	✓	-	-	-	-	-	-	-	-	-
	SAKF	-	-	✓	-	-	-	✓	-	-	-	-
	Random Forest	-	-	-	✓	-	-	-	-	-	-	-
	CNN	✓	-	-	-	✓	-	-	-	✓	-	-
	NB	-	-	-	-	-	✓	-	-	-	✓	-
	SVM	-	-	-	-	-	✓	-	-	-	-	-
	Image Hashing	-	-	-	-	-	-	-	✓	-	-	-
	BN	-	-	-	-	-	-	-	-	-	✓	-
	j48	-	-	-	-	-	-	-	-	-	✓	-
	jRip	-	-	-	-	-	-	-	-	-	✓	-
	kNN	-	-	-	-	-	-	-	-	-	-	✓
Datasets	[29], [30]	✓	-	-	-	-	-	-	-	-	-	-
	Self Constructed	-	✓	-	-	-	-	-	-	-	✓	✓
	Tor Images [26]	-	-	✓	-	-	-	-	-	-	-	-
	SURFnet	-	-	-	✓	-	-	-	-	-	-	-
	ImageCLEF [31]	-	-	-	-	✓	-	-	-	-	-	-
	DUTA-10K [32]	-	-	-	-	-	✓	-	-	-	-	-
	TOIC [33]	-	-	-	-	-	-	✓	-	-	-	-
	USC-SIPI [34]	-	-	-	-	-	-	-	✓	-	-	-
	[35]	-	-	-	-	-	-	-	-	✓	-	-

researchers proposed a machine learning-based framework based on supervised (CNN) and unsupervised (LDA) learning techniques to infer the massive dataset containing images to encounter this issue. Thus, this method stands out to be more proficient in investigating the material being transmitted on darknet markets. The authors in [18] proposed a natural language processing (NLP) based system for detecting the legal and illegal communications being carried on the Darknet. The researchers claim to fulfill the gaps that are present in the current NLP-based models. The model detects the legal or illegal text on the Darknet by comparing websites with similar content. The model takes a drug-relevant website as a test case. The model compares text and linguistic used on both websites and then distinguishes them based on POS tags and the availability of their termed objects in Wikipedia.

Cybercrimes, illegal drugs dealing, and cryptocurrency markets are top trending over the Darknet recently. The authors study and identify the research done over darknet technologies and specifications to highlight the gaps and difficulties in exploiting the illegal activities going over the Darknet. The researchers in [36] discussed the limitations and difficulties faced while identifying the criminal activities over the cyber network as the Darknet is protected with specific network encryptions and configurations are hidden from search engines. This not only helps government agencies pursue the culprits, but it will also help keep peace in society. The researchers also focused on the internet for crimes, exploiting Tor-services users' identity and drug dealing, and proposed a system to inspect and explore unidentified online illegitimate markets. The authors in [37] review the latest technologies and researches that encounter the vulgar activities going on the Darknet, which includes money laundry, child pornography, and illegitimate drug trafficking. However, cryptocurrency is required for transactions over the Darknet, which includes bitcoins. The authors review and highlight the latest studies that deanonymize the culprits and their activities over the Darknet, resulting in aid for the cybercrime investigation agencies.

The authors in [38] performed a Tor network in-depth review. The authors mention all the darknet pros and cons and the nature of darknet network traffic flow. The anonymization of the Darknet is considered one of the essential features since it allows users to anonymously perform any activity, whether it is legal or not. The authors also review the issues of anonymous payments with cryptocurrency. The authors conclude this extensive review by enlightening the future aspects and challenges of the Darknet. The researchers in [26] proposed a SAKF system that classifies the images by analyzing only pertinent portions of the image compounding salience maps such that it chooses only those areas of the image which have more noticeable information with BoVW. This system is tested on seven different publicly available datasets on which the system showed an accuracy gain of 1.64 to 15.73 higher than baseline approaches which used BoVW using dense SIFT (Scale-Invariant Feature Transform) descriptors.

The researchers in [27] proposed a programmed model that recognizes the activities being done on the Tor network by analyzing the activities' snapshots. The researchers in [12] combined three proficient algorithms with the help of a rectification network to enhance the algorithm's text recognition procedure. The authors evaluated this composition of algorithms on four text datasets and TOICO-1K image datasets of Tor for analyzing the text. The proposed model shows the highest proficiency over the ICDAR 2015 dataset.

The researchers in [11], [39], [40] proposed a framework based on network flow features for Tor traffic analysis and multi-level cataloging. The model can detect the anonymous traffic L1, L2, and L3 on multiple platforms, including mobile and PC. The author claim that the impact of time-related features is higher than that of the non-time-related features on the mobile platform, while it is the opposite on the PC platform. The researcher in [28] proposed a machine learning approach

to text classification over the Russian language-based Tor sites. The proposed model analyzes the text and categorize it for the forensics purpose. For a quick analysis, we conclude state-of-the-art studies in Table 1.

III. DATASET SELECTION

Other datasets exist, such as Anon17 and Darknet Usage Text Address (DUTA)-10K-GVIS Lab (2019). Anon17 dataset consists of three anonymity tools: Tor, I2P, and JonDonym [13]. DUTA dataset contains 25 categories of legal and illegal activities with over 10,367 manually labeled onion domains.[1] We select recently published dataset [15] that is an amalgamated of previously two identical datasets [30], and [29] with the enhancement of VPN and Tor-traffic categories. This dataset contains 85 features, including 2 labels, traffic_type: Tor, Non-Tor, VPN, Non-VPN and app_category: Audio-Stream, Browsing, Chat, E-mail, P2P, Transfer, Video-Stream, VOIP. Table 2 shows the traffic categories and number of samples of each traffic type. Table 3 shows each app category with the number of samples.

TABLE 2. Traffic category.

Traffic Category	Samples
Non-Tor	93357
Tor	1393
Non-VPN	23864
VPN	22920

TABLE 3. Application category.

App Category	Samples
Audio Streaming	18065
Browsing	32809
Chat	11479
Email	6146
File Transfer	11183
P2P	48521
Video Streaming	9768
VOIP	3567

IV. DARKDETECT

Figure 1 represents the proposed approach, starting with feature extraction from the data lake, and finding out the essential features for detecting each traffic category. From the table 2 and 3, we came to the point that data is an imbalance in each traffic category and app category. Data is balanced through the oversampling technique. Next, at the classification level, we detect traffic types and categories their applications through different ML and DL techniques explained in section IV-D.

A. HANDLING MISSING DATA

This section describes various ways to handle missing data. Missing data handling is a critically essential task to improve the machine effectively, and deep learning classifier's performance [41]. Missing data refers to the data that is not available (Null) or inapplicable or infinite data or the data collection event that never happened. Since the dataset is vast, we remove the rows that contain missing information. It will not have any effect on the learning process [42]. Table 4 provides the samples of missing data highlighted in bold text. This data can lead to misleading results of the machine and deep learning classifier.

TABLE 4. Missing data in the dataset.

Bwd Packet Length Std	Flow Bytes/s	Flow Packets/s	Flow IAT Mean	Flow IAT Std
0	-	3273.322	611	**NaN**
0	0	1904.762	1050	-
NaN	**NaN**	**Infinity**	0	0
0	536.9445	**Infinity**	186239	0
0	0	5494.505	**Infinity**	0
NaN	**NaN**	3.0371	658523	**NaN**

B. DATA BALANCING

The dataset balancing is one of the vital parts of data prepossessing. It optimizes the dataset so that every class present in the dataset may get recognized to be processed and predicted accurately. According to the dataset structure used for this approach, there is high unbalancing between classes instance, leading to significantly less F1-Score but higher accuracy and precision. We use SMOTE [8], [43] for data balancing. SMOTE over-samples the instances of the majority class, meanwhile keeping the original representation of the dataset. We increase the Tor class by 20% since Tor was the only class with the least instances.

C. FEATURE SELECTION

We apply the Principle Component Analysis (PCA) [6], DT [6], [44], and XGB [45] for feature selection to extract 20 important features. A DT consists of roots, branches, nodes, and leaves. One branch consists of nodes and comprises one feature. The occurrence of each feature from root to leaves depicts the importance of each feature. Similarly, the XGB classifier selects the optimal features using feature importance criteria. It selects the feature based on a threshold (i.e., 0.08). If the threshold is too low, no feature will be selected.

We compare the selected 20 essential features with the performance of 22 selected important features of the state-of-the-art study as shown in Table 5. The state-of-the-art study uses 15 features in the classification step from these 22 features, while we use 20 features extracted through the below feature selection techniques.

Principal component analysis (PCA) is an unsupervised algorithm that produced linear combinations of the actual features [6]. The new features are orthogonal, which means that they are not correlated with each other. Furthermore, those features are ranked in order of their explained variance.

[1] http://gvis.unileon.es/dataset/duta-darknet-usage-text-addresses-10k/

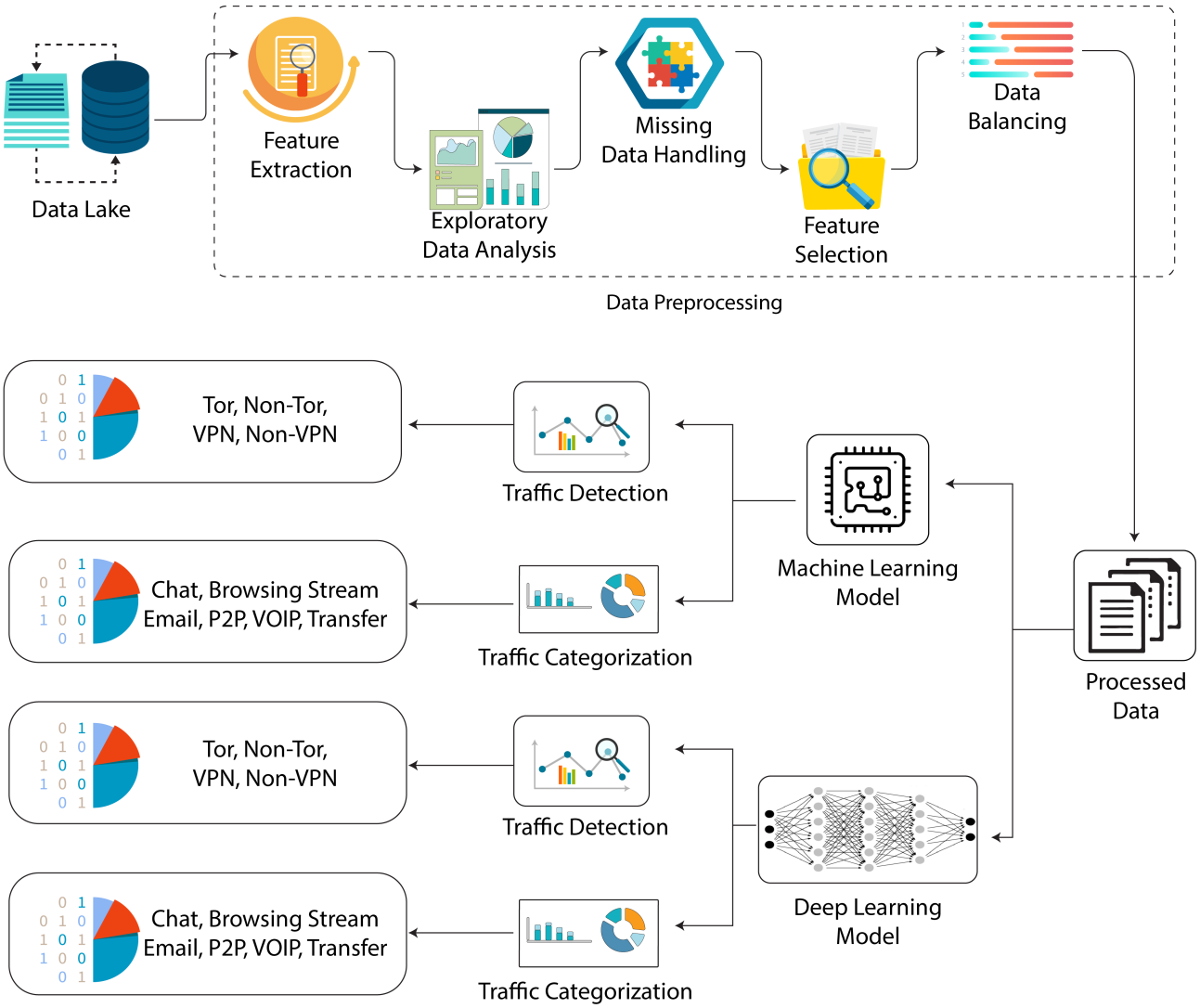


FIGURE 1. Taxonomy of the proposed approach.

Below are the machine learning classifiers that we used to select the features based on their importance according to accuracy.

Decision Tree Classifier works for both regression as well as classification problems [6], [44], [46]. The decision is simple to understand, interpret and visualize and can handle un-prepared data. Nonlinear parameters do not affect the performance of the DT. The performance of the decision can be measured using entropy (unpredictability) and information gain. The less the entropy and highest information, the classifier is more accurate.

XGBClassifier is a decision-tree-based ensemble ML algorithm that uses a gradient-boosting framework [45]. XGB improves upon the base Gradient Boosting Machines (GB) framework through algorithmic enhancements and systems optimization. The core XGB algorithm is parallelizable, which means it does parallelization within a single DT. XGB performs tree pruning, the stopping criteria for tree splitting within the GB Machines framework, which is greedy and depends upon the negative loss criterion at the split point. XGB admits sparse features for inputs by automatically learning the best missing value depending upon training loss and efficiently handling various sparsity patterns in the data. This approach supports both regression and classification predictive modeling problems.

D. CLASSIFICATION

The dataset contains traffic types such as Tor, Non-Tor, VPN, Non-VPN, and app_categories such as Audio-Stream, Browsing, Chat, E-mail, P2P, Transfer, Video-Stream,

TABLE 5. Important feature selected by the proposed approach.

Rank	Selected Features
1	Idle Max
2	FWD Init Win Bytes
3	Idle Mean
4	Idle Min
5	Fwd Seg Size Min
6	Subflow Fwd Packets
7	Flow Duration
8	Flow IAT Max
9	Flow IAT Min
10	Flow IAT Mean
11	Fwd Packets/s
12	Flow Packets/s
13	Bwd Init Win Bytes
14	Protocol
15	FIN Flag Count
16	Bwd Packets/s
17	Fwd IAT Max
18	Bwd Packet Length Mean
19	Bwd Packet Length Min
20	Fwd IAT Total

VOIP. We apply modified deep learning techniques: Convolution-Long Short Term Memory (CNN-LSTM) and Convolution-Gated Recurrent Unit (CNN-GRU) to recognize traffic types. We also analyze the recognition of the traffic categories using the machine learning techniques: GB, DT, RFR, and XGB classifiers. We tune and select the parameters based on the efficiency of the model (i.e., F1-score). We select the combination that was providing us best results. In the results, we show the table with all combinations and conclude the best results.

Gradient Boosting Classifier (GB) produces a prediction model in the form of a set of weak classifier models, like decision trees. Successive decision trees are generated during the learning/training process. The algorithm builds the 1st model to predict the value and calculate the loss, the difference between the 1st model's result and the true value. Moreover, after the first step, the second classifier model is then built to predict the loss. Until a satisfactory result is achieved, this whole process continues. The GB is to find new trees that minimize the loss iteratively. The loss function is used to measure how many errors are made by the classifier.

Random Forest Regressor (RFR) is an ensemble ML technique proficient in classification and regression tasks using multiple DT and bagging, which is a statistical technique. Bagging and boosting algorithm, these 2 are the most well-known ensemble techniques that aim to tackle high bias and high variance. An Rf is not just averaging the prediction of decision trees; it uses two main vital concepts that give it title random: (1) while building decision trees, random sampling of training was observed, (2) random subgroups features for splitting nodes. Rf builds multiple decision trees and aggregates their predictions to get more stable and accurate results rather than relying on individual decision trees. Entropy uses the probability of a particular result to decide how the tree nodes should branch. En is a more mathematical-intensive cause of logarithmic function, which is used to calculating it.

CNN-LSTM performs multiple tasks to deliver a proficient accuracy. CNN layer extracts the features of input data, and LSTM supports the sequence prediction. The CNN-LSTM model is mainly utilized for data input in series, just as in continuous network traffic flow.

$$h_t = H(W_{hx}x_t + W_{hh}h_{t-1} + b_h) \tag{1}$$
$$p_t = W_{hy}y_{t-1} + b_y \tag{2}$$

The Equations 1 and 2 represents main computing equations where series input is denoted by x_t, series output is denoted by y_t, hidden memory cells are denoted by h_t, weight matrices are denoted by W, and bias vectors are denoted by b. Following equations computer the hidden state of memory cells:

$$i_t = \sigma(W_{ix}x_t + W_{hh}h_{t-1} + W_{ic}c_{t-1} + b_i) \tag{3}$$
$$f_t = \sigma(W_{fx}x_t + W_{hh}h_{t-1} + W_{fc}c_{t-1} + b_f) \tag{4}$$
$$c_t = f_t * c_{t-1} + i_t * g(W_{cx}x_t + W_{hh}h_{t-1} + W_{cc}c_{t-1} + b_c) \tag{5}$$
$$o_t = \sigma(W_{ox}x_t + W_{hh}h_{t-1} + W_{oc}c_{t-1} + b_o) \tag{6}$$
$$h_t = o_t * h(c_t) \tag{7}$$

CNN layer extracts the features of input data, and GRU supports the sequence prediction. **CNN-GRU** is an improvement in the basic functionality of the standard recurrent neural network. Basic RNN faces a vanishing gradient problem that GRU encounters by updating and resetting the gate. GRU also maintains the ability to maintain the information for a long time or discard irrelevant information to make efficiently accurate predictions.

$$z_t = \delta(W^{(t)}x_t + U^z h_{t-1}) \tag{8}$$

The Equation 8 represents the gate updation process in which x_t is inserted into the network unit, $W(z)$ represents it's own weight. $h_(t-1)$ contains the information for the previous unit and gets multiplied by own weight $U(z)$.

$$r_t = \delta(W^{(r)}x_t + U^r h_{t-1}) \tag{9}$$

The Equation 9 represents the reset part functionality of GRU.

$$h_t = z_t \odot h_{h-1} + (1 - z_t) \odot h'_t \tag{10}$$

The Equation 10 represents the final memory at a particular time step.

V. EVALUATION AND RESULTS

We use Precision, Recall, F-score, and Accuracy evaluation metrics to compare the proposed approach's performance. Table 6 shows the parameters with tuned values of modified convolution-LSTM to achieve the best accuracy and loss.

TABLE 6. Tuned parameters.

Parameter	Value
Hidden layers Function	RELU
Output_layer Function	Softmax
Loss Function	Categorical
Optimizer	Adam
Epoch	180
Batch Size	46
Estimators	200
Maximum Depth	9
Early Stopping Monitor	Best weights

A. HARDWARE ENVIRONMENT

The experiment's computing environment is set as Intel(R) Corei7, 8th Generation with 16 GB RAM, Windows 10 OS, and Python version 3.7.6. For parameter tuning, we used GPUs from Google Colab and Tesla lab. Table 7 shows the computing environment in which the experiment is conducted.

TABLE 7. Computing environment.

Environment	Parameter
Operating System	Windows
CPU	8th Gen-intel Corei7
RAM	12GB
Python Version	3.7.6

Results are extracted based on feature selection approaches in combination with classification techniques. For feature selection, we use PCA, DT, and XGB. For classification, we modify the CNN model with a hidden layer of LSTM and the CNN model with a hidden layer of GRU.

TABLE 8. Performance evaluation metrics shows comparison of classification techniques. The Precision, Recall, and F-score are in the range of [0-1] with 1 being the highest. The highest values are in bold.

Feature Selection	Model	Precision	Recall	F1 Score
XGB	CNN-GRU	0.96	0.91	0.93
	CNN-LSTM	**0.97**	**0.95**	**0.96**

B. DARKNET TRAFFIC DETECTION

We extensively analyze traffic categorization and conclude that the feature selection approach XGB selects the best feature than DT and PCA and improves accuracy with the CNN-LSTM classifier. We apply the same hierarchy for traffic detection and present the best results in Table 8. As shown in Table 8, we achieve 1%, 4%, and 3% better Precision, recall, and F-score through CNN-LSTM than CNN-GRU on traffic Tor, Non-Tor, VPN, and Non-VPN. Figure 2 shows the results of our technique.

C. DARKNET TRAFFIC CATEGORIZATION

Table 10, 9 and table 11 represents results on layer2 with traffic categories Audio-Stream, Chat, Email, P2P, VOIP, Video-Stream, Transfer, and Browsing. In darknet traffic categorization, we first analyze the performance of machine learning classifiers. The results of ML classifiers are based on 20 best-selected features and after data balancing through SMOTE. Table 9 shows that RFR achieves 71% F-score while GB achieves 81% F-score, which is 10% higher than RFR. The DT achieves 84% F-score, which is 3% and 13% higher than GB and RFR. In comparison, XGB achieves 85% F-score, which is 1%, 4%, and 14% higher than DT, GB, and RFR. Although ML models show significant performance on darknet traffic categorization, categorization still needs some improvement.

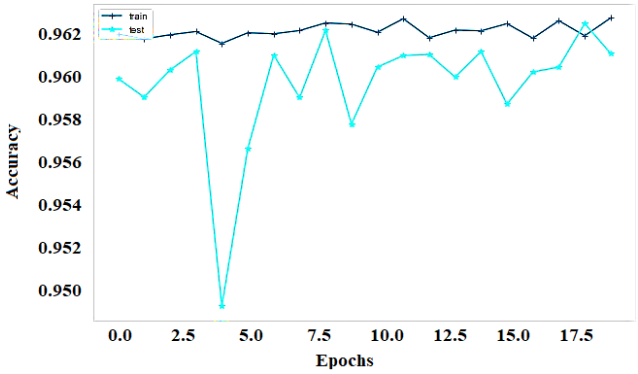


FIGURE 2. Accuracy curves of the proposed DarkDetect approach.

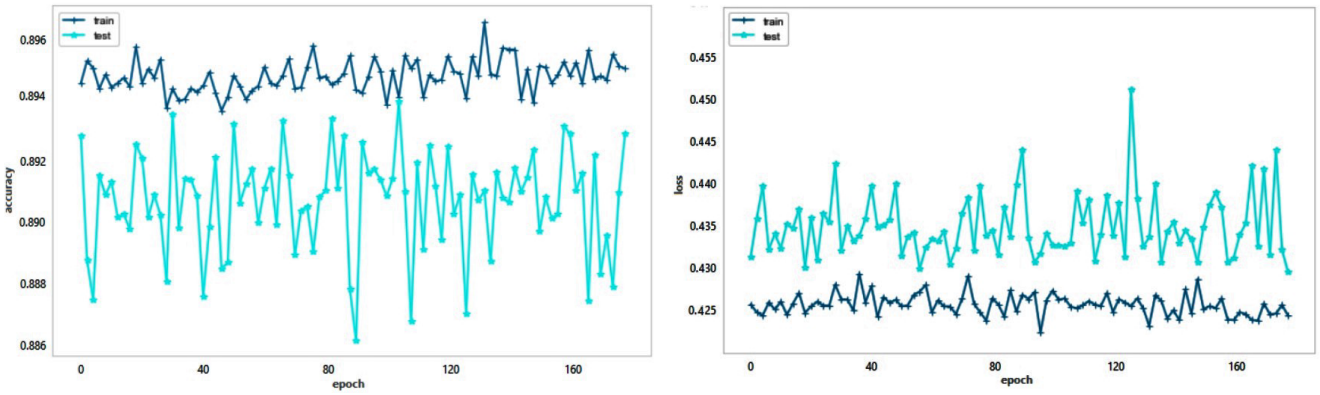
TABLE 9. Performance evaluation metrics shows comparison of classification techniques. The Precision, Recall, and F-score are in the range of [0-1] with 1 being the highest. The highest values are in bold.

	Model	Precision	Recall	F1 Score
Data Balancing	DT	0.85	0.84	0.84
	GB	0.83	0.81	0.81
	RFR	0.71	0.72	0.71
	XGB	**0.85**	**0.85**	**0.85**

TABLE 10. Performance evaluation metrics show a comparison of feature selection and classification techniques as a combination. The Precision, Recall, and F-score is in the range of [0-1], with 1 being the highest. The highest values are in bold.

Feature Selection	Classification	Precision	Recall	F1 Score	AUC
PCA	CNN-LSTM	0.70	0.73	0.72	0.81
	CNN-GRU	0.75	0.73	0.74	0.87
DT	CNN-LSTM	0.85	0.84	0.84	0.90
	CNN-GRU	0.79	0.80	0.80	0.87
XGB	CNN-GRU	0.81	0.80	0.82	0.89
	CNN-LSTM	**0.90**	**0.88**	**0.89**	**0.95**

Table 10 shows that CNN-GRU performs 2% better than CNN-LSTM on features selected by PCA. In comparison, CNN-LSTM performs 4% better than CNN-GRU on features that DT selects. Meanwhile, in conclusion, the CNN-LSTM model significantly improves 9% than CNN-GRU on features selected by XGB. So, in table 11 we compare our best results extracted through CNN-LSTM on features which are selected by XGB with the state-of-the-art study [15].



(a) Accuracy Curve (b) Loss curve

FIGURE 3. Accuracy and Loss curves of DarkDetect.

Figure 3 shows the results of our technique. It is concluded that we achieve overall 3% better accuracy, while with the loss of 0.43, which is 0.07 lower than the study [15].

TABLE 11. Comparison results on evaluation metrics of proposed approach and state-of-the-art study [15]. Key: Precision-p, F-score-F1, Recall-R.

Categories	Lashkari et al. [15]			DarkDetect		
	P	R	F1	P	R	F1
Browsing	0.55	0.47	0.51	0.82	0.89	**0.85**
Chat	0.90	0.86	0.88	0.86	0.81	0.83
Email	0.66	0.67	0.67	0.85	0.84	**0.84**
File-Transfer	0.74	0.75	0.75	0.88	0.83	**0.85**
P2P	0.90	0.95	0.93	0.95	0.93	**0.94**
Video Streaming	0.82	0.88	0.85	0.87	0.86	**0.86**
Audio Streaming	0.92	0.92	0.92	0.83	0.81	0.82
VOIP	0.58	0.61	0.59	0.89	0.87	**0.87**
Average Accuracy	86%			89%		

As shown in Table 11, we achieve 44%, 17%, 10% 1%, 1%, and 27% better F-score on Browsing, File-Transfer, P2P, Video-Streaming, and VOIP traffic categories than base paper. This is because of 20 important selected features. Meanwhile, we got 5% and 10% fewer F-score on Chat and Audio-Streaming traffic categories than base paper. This comparative analysis proves that the proposed approach XGB-CNN-LSTM increased 73% F-score on 6 traffic categories while the minimal loss of only 15% than the state-of-the-art study [15].

Finally, after a comprehensive analysis and experiment, this paper answers the identified research question. The answer to *RQ1* would be that darknet traffic can be detected using the proposed deep detect pipeline, which consists of XGB based feature selection, data balancing using SMOTE, and classification using CNN-LSTM. The answer to *RQ2* would be that XGB feature selection along with CNN-LSTM affects positively predicts the darknet traffic. *RQ3* can be answered since the minority class instances of Tor traffic are being misclassified in the testing process due to inadequate training. SMOTE improves the representation of Tor class and classification accuracy.

VI. CONCLUSION

In this paper, an approach is proposed to detect and categorize darknet traffic. The proposed approach consists of advanced fine-tuned machine learning algorithms and a convolutional neural network-based deep learning classifier. These classifiers are tested for the classification over state-of-the-art [30] dataset, which contains 8 categories of network traffic packets. Among the tested machine learning classifiers, the classifier XGB stands out to be the most proficient than other competitors by yielding a proficient average F-score of 85%. Moreover, the modified classifier CNN-LSTM and the XGB feature selection approach outperform the state-of-the-art study [15] with a gain of 3% accuracy on average. In the future, it is intended to extend this work by testing and combining the classifiers in an ensemble approach and further creating a personalized generated dataset that is being loaded with more precise information that is required to train a classifier with more proficiency.

REFERENCES

[1] M. Mittal, C. Iwendi, S. Khan, and A. R. Javed, "Analysis of security and energy efficiency for shortest route discovery in low-energy adaptive clustering hierarchy protocol using Levenberg-Marquardt neural network and gated recurrent unit for intrusion detection system," *Trans. Emerg. Telecommun. Technol.*, vol. 32, no. 6, p. e3997, Jun. 2021.

[2] W. Ahmed, F. Shahzad, A. R. Javed, F. Iqbal, and L. Ali, "WhatsApp network forensics: Discovering the IP addresses of suspects," in *Proc. 11th IFIP Int. Conf. New Technol., Mobility Secur. (NTMS)*, Apr. 2021, pp. 1–7.

[3] C. Iwendi, S. U. Rehman, A. R. Javed, S. Khan, and G. Srivastava, "Sustainable security for the Internet of Things using artificial intelligence architectures," *ACM Trans. Internet Technol.*, vol. 21, no. 3, pp. 1–22, Jun. 2021.

[4] K. Misata, "The Tor project: An inside view," *XRDS, Crossroads, ACM Mag. Students*, vol. 20, no. 1, pp. 45–47, Sep. 2013.

[5] F. Soro, M. Allegretta, M. Mellia, I. Drago, and L. M. Bertholdo, "Sensing the noise: Uncovering communities in darknet traffic," in *Proc. Medit. Commun. Comput. Netw. Conf. (MedComNet)*, Jun. 2020, pp. 1–8.

[6] A. R. Javed, Z. Jalil, S. A. Moqurrab, S. Abbas, and X. Liu, "Ensemble AdaBoost classifier for accurate and fast detection of botnet attacks in connected vehicles," *Trans. Emerg. Telecommun. Technol.*, p. e4088, Aug. 2020.

[7] C. Iwendi, Z. Jalil, A. R. Javed, T. Reddy G., R. Kaluri, G. Srivastava, and O. Jo, "KeySplitWatermark: Zero watermarking algorithm for software protection against cyber-attacks," *IEEE Access*, vol. 8, pp. 72650–72660, 2020.

[8] S. U. Rehman, M. Khaliq, S. I. Imtiaz, A. Rasool, M. Shafiq, A. R. Javed, Z. Jalil, and A. K. Bashir, "DIDDOS: An approach for detection and identification of distributed denial of service (DDoS) cyberattacks using gated recurrent units (GRU)," *Future Gener. Comput. Syst.*, vol. 118, pp. 453–466, May 2021.

[9] S. I. Imtiaz, S. U. Rehman, A. R. Javed, Z. Jalil, X. Liu, and W. S. Alnumay, "DeepAMD: Detection and identification of Android malware using high-efficient deep artificial neural network," *Future Gener. Comput. Syst.*, vol. 115, pp. 844–856, Feb. 2021.

[10] S. Kumar, H. Vranken, J. V. Dijk, and T. Hamalainen, "Deep in the dark: A novel threat detection system using darknet traffic," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Dec. 2019, pp. 4273–4279.

[11] L. Wang, H. Mei, and V. S. Sheng, "Multilevel identification and classification analysis of Tor on mobile and PC platforms," *IEEE Trans. Ind. Informat.*, vol. 17, no. 2, pp. 1079–1088, Feb. 2021.

[12] P. Blanco-Medina, E. Fidalgo, E. Alegre, and F. Jánez-Martino, "Improving text recognition in Tor darknet with rectification and super-resolution techniques," in *Proc. 9th Int. Conf. Imag. Crime Detection Prevention*, 2019, pp. 32–37.

[13] A. Montieri, D. Ciuonzo, G. Aceto, and A. Pescapé, "Anonymity services Tor, I2P, JonDonym: Classifying in the dark (web)," *IEEE Trans. Dependable Secure Comput.*, vol. 17, no. 3, pp. 662–675, May 2020.

[14] A. Montieri, D. Ciuonzo, G. Bovenzi, V. Persico, and A. Pescapé, "A dive into the dark web: Hierarchical traffic classification of anonymity tools," *IEEE Trans. Netw. Sci. Eng.*, vol. 7, no. 3, pp. 1043–1054, Mar. 2019.

[15] A. H. Lashkari, G. Kaur, and A. Rahali, "DIDarknet: A contemporary approach to detect and characterize the darknet traffic using deep image learning," in *Proc. 10th Int. Conf. Commun. Netw. Secur.*, Nov. 2020, pp. 1–13.

[16] E. F. Fernandez, R. A. V. Carofilis, F. J. Martino, and P. B. Medina, "Classifying suspicious content in Tor Darknet," 2020, *arXiv:2005.10086*. [Online]. Available: http://arxiv.org/abs/2005.10086

[17] A. Berman and C. L. Paul, "Making sense of Darknet markets: Automatic inference of semantic classifications from unconventional multimedia datasets," in *Proc. Int. Conf. Hum.-Comput. Interact.* Cham, Switzerland: Springer, 2019, pp. 230–248.

[18] L. Choshen, D. Eldad, D. Hershcovich, E. Sulem, and O. Abend, "The language of legal and illegal activity on the Darknet," 2019, *arXiv:1905.05543*. [Online]. Available: http://arxiv.org/abs/1905.05543

[19] A. Basit, M. Zafar, X. Liu, A. R. Javed, Z. Jalil, and K. Kifayat, "A comprehensive survey of AI-enabled phishing attacks detection techniques," *Telecommun. Syst.*, vol. 76, no. 1, pp. 1–16, 2020.

[20] A. Basit, M. Zafar, A. R. Javed, and Z. Jalil, "A novel ensemble machine learning method to detect phishing attack," in *Proc. IEEE 23rd Int. Multitopic Conf. (INMIC)*, Nov. 2020, pp. 1–5.

[21] C. Mo, W. Xiaojuan, H. Mingshu, J. Lei, K. Javeed, and X. Wang, "A network traffic classification model based on metric learning," *Comput., Mater. Continua*, vol. 64, no. 2, pp. 941–959, 2020.

[22] B. Xiong, K. Yang, J. Y. Zhao, and K. Q. Li, "Robust dynamic network traffic partitioning against malicious attacks," *J. Netw. Comput. Appl.*, vol. 87, pp. 20–31, Jun. 2017.

[23] C. Du, S. Liu, L. Si, Y. Guo, and T. Jin, "Using object detection network for malware detection and identification in network traffic packets," *Comput., Mater. Continua*, vol. 64, no. 3, pp. 1785–1796, 2020.

[24] W. Tian, X. Ji, W. Liu, G. Liu, R. Lin, J. Zhai, and Y. Dai, "Defense strategies against network attacks in cyber-physical systems with analysis cost constraint based on honeypot game model," *Comput., Mater. Continua*, vol. 60, no. 1, pp. 193–211, 2019.

[25] S. Lagraa, Y. Chen, and J. François, "Deep mining port scans from darknet," *Int. J. Netw. Manage.*, vol. 29, no. 3, p. e2065, May 2019.

[26] E. Fidalgo, E. Alegre, L. Fernández-Robles, and V. González-Castro, "Classifying suspicious content in Tor darknet through semantic attention keypoint filtering," *Digit. Invest.*, vol. 30, pp. 12–22, Sep. 2019.

[27] R. Biswas, V. González-Castro, E. Fidalgo, and E. Alegre, "Perceptual image hashing based on frequency dominant neighborhood structure applied to Tor domains recognition," *Neurocomputing*, vol. 383, pp. 24–38, Mar. 2020.

[28] I. D. Buldin and N. S. Ivanov, "Text classification of illegal activities on onion sites," in *Proc. IEEE Conf. Russian Young Researchers Electr. Electron. Eng. (EIConRus)*, Jan. 2020, pp. 245–247.

[29] G. Draper-Gil, A. H. Lashkari, M. S. I. Mamun, and A. A. Ghorbani, "Characterization of encrypted and VPN traffic using time-related features," in *Proc. 2nd Int. Conf. Inf. Syst. Secur. Privacy*, 2016, pp. 407–414.

[30] A. H. Lashkari, G. Draper-Gil, M. S. I. Mamun, and A. A. Ghorbani, "Characterization of Tor traffic using time based features," in *Proc. ICISSp*, 2017, pp. 253–262.

[31] T. Tsikrika, A. Popescu, and J. Kludas, "Overview of the Wikipedia image retrieval task at ImageCLEF 2011," in *Proc. CLEF, Notebook Papers/Labs/Workshop*, vol. 4, 2011, p. 5.

[32] M. W. Al-Nabki, E. Fidalgo, E. Alegre, and L. Fernández-Robles, "ToRank: Identifying the most influential suspicious domains in the Tor network," *Expert Syst. Appl.*, vol. 123, pp. 212–226, Jun. 2019.

[33] E. Fidalgo, E. Alegre, V. González-Castro, and L. Fernández-Robles, "Boosting image classification through semantic attention filtering strategies," *Pattern Recognit. Lett.*, vol. 112, pp. 176–183, Sep. 2018.

[34] A. G. Weber, "The USC-SIPI image database version 5," *USC-SIPI Rep.*, vol. 315, no. 1, 1997.

[35] P. B. Medina, E. F. Fernández, E. A. Gutiérrez, and M. W. Al Nabki, "Detecting textual information in images from onion domains using text spotting," in *Proc. 39th Jornadas de Automática, Actas*, Badajoz, Spain, 2018, pp. 975–982.

[36] V. Adewopo, B. Gonen, S. Varlioglu, and M. Ozer, "Plunge into the underworld: A survey on emergence of darknet," in *Proc. Int. Conf. Comput. Sci. Comput. Intell. (CSCI)*, Dec. 2019, pp. 155–159.

[37] W. Han, V. Duong, L. Nguyen, and C. Mier, "Darknet and bitcoin deanonymization: Emerging development," in *Proc. Zooming Innov. Consum. Technol. Conf. (ZINC)*, May 2020, pp. 222–226.

[38] A. Nastuła, "Dilemmas related to the functioning and growth of Darknet and the onion router network," *J. Sci. Papers Social Develop. Security*, vol. 10, no. 2, pp. 3–10, Apr. 2020.

[39] K. Shahbar and A. N. Zincir-Heywood, "Traffic flow analysis of Tor pluggable transports," in *Proc. 11th Int. Conf. Netw. Service Manage. (CNSM)*, Nov. 2015, pp. 178–181.

[40] K. Shahbar and A. N. Zincir-Heywood, "Benchmarking two techniques for Tor classification: Flow level and circuit level classification," in *Proc. IEEE Symp. Comput. Intell. Cyber Secur. (CICS)*, Dec. 2014, pp. 1–8.

[41] A. R. Javed, S. U. Rehman, M. U. Khan, M. Alazab, and H. U. Khan, "Betalogger: Smartphone sensor-based side-channel attack detection and text inference using language modeling and dense MultiLayer neural network," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 20, no. 5, pp. 1–17, Jun. 2021.

[42] M. Asad, M. Asim, T. Javed, M. O. Beg, H. Mujtaba, and S. Abbas, "DeepDetect: Detection of distributed denial of service attacks using deep learning," *Comput. J.*, vol. 63, no. 7, pp. 983–994, Jul. 2020.

[43] R. Blagus and L. Lusa, "Class prediction for high-dimensional class-imbalanced data," *BMC Bioinf.*, vol. 11, no. 1, pp. 1–17, Dec. 2010.

[44] Y. P. P. A. F. Pavel and B. C. Soares, "Decision tree-based data characterization for meta-learning," in *Proc. IDDM-2002*, vol. 111, 2002.

[45] A. Abbasi, A. R. Javed, C. Chakraborty, J. Nebhen, W. Zehra, and Z. Jalil, "ElStream: An ensemble learning approach for concept drift detection in dynamic social big data stream learning," *IEEE Access*, vol. 9, pp. 66408–66419, 2021.

[46] A. R. Javed, T. Baker, M. Asim, M. Beg, and A. H. Al-Bayatti, "AlphaLogger: Detecting motion-based side-channel attack using smartphone keystrokes," Tech. Rep., 2020.