IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Edge Computing Intelligence Using Robust Feature Selection for Network Traffic Classification in Internet-of-Things

BUSHRA MOHAMMED[1], MOSAB HAMDAN[2] (Member, IEEE), SB JOSEPH[3], HA JAMIL[4], SULEMAN KHAN[5], ABDALLAH ELHIGAZI [6], DANDA B. RAWAT[7] (Senior Member, IEEE), ISMAHANI ISMAIL[2], and MN MARSONO[2] (Member, IEEE)

[1]Faculty of Computer and Statistics Studies, University of Kordofan, Sudan
[2]School of Electrical Engineering, Faculty of Engineering, University Teknology Malaysia
[3]Department of Computer Engineering, Faculty of Engineering, University of Maiduguri,Borno State Nigeria
[4]Faculty of Computer Science and IT Engineering, University of Elimam Elmahdi Kosti, Sudan
[5]Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne NE1 8ST, U.K
[6]School of Computing, Faculty of Engineering, University Teknology Malaysia
[7]Department of Electrical Engineering & Computer Science, Howard University, Washington, DC, USA

Corresponding authors: Bushra Mohammed (e-mail: bushra091215@gmail.com), MN Marsono (e-mail: mnadzir@utm.my)

ABSTRACT Internet-of-Things (IoT) devices are massively interconnected, which generates a massive amount of network traffic. The concept of edge computing brings a new paradigm to monitor and manage network traffic at the network's edge. Network traffic classification is a critical task to monitor and identify Internet traffic. Recent traffic classification works suggested using statistical flow features to classify network traffic accurately using machine learning techniques. The selected classification features must be stable and can work across different spatial and temporal heterogeneity. This paper proposes a feature selection mechanism called Ensemble Weight Approach (EWA) for selecting significant features for Internet traffic classification based on multi-criterion ranking and selection mechanisms. Extensive simulations have been conducted using publicly-available traces from the University of Cambridge. The simulation results demonstrate that EWA is capable of identifying stable features subset for Internet traffic identification. EWA-selected features improve the mean accuracy up to 1.3% and reduce RMSE using fewer features than other feature selection methods. The smaller number of features directly contributes to shorter classification time. Furthermore, the selected features can train stable traffic classification generative models irrespective of the dataset's spatial and temporal differences, with consistent accuracy up to 97%. The overall performance indicates that EWA-selected statistical flow features can improve the overall traffic classification.

INDEX TERMS Edge computing, Internet-of-Things, Network traffic classification, Feature selection

I. INTRODUCTION

THE introduction of the Internet-of-Things (IoT) has benefited numerous sectors like healthcare, manufacturing, finance, and entertainment. The massive IoT devices' interconnectivity raises serious concerns since it resulted in high network traffic. Monitoring and managing network traffic, especially at the network's edge, requires accurate and efficient network traffic classification. One of the factors for efficient and accurate network traffic classification is the selected classification features that are stable and can work across different spatial and temporal heterogeneity.

Traffic application identification is a fundamental and critical task in network traffic management [1]. The limitation of port-based e.g. [2]–[4] and payload-based strategies e.g. [3], [5]–[8] prompts the use of statistical flow features e.g. [9], [10] for traffic classification. The latter provides the pliability to identify network traffic in contrast to port-based and signature-based strategies since this type of traffic identifier

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037492, IEEE Access

is not affected by detection avoidance mechanisms such as dynamic port numbers and payload encryption.

Identifying the classes of Internet traffic using statistical flow features is non-trivial because of the high dimensionality of traffic features used for traffic classification. Preferably, the usage of many features would boost the ability to differentiate Internet traffic [11], [12]. Nonetheless, it is not always so in practice because not every feature is informative and useful. Some statistical flow features may not be relevant and uninformative, while others may have high inter-correlation with each other features and thus redundant [13], [14]. The use of less significant traffic features affects the efficiency and accuracy of network traffic classification [13], [15]–[17].

Several feature selection (FS) techniques have been proposed in literature [15], [17]–[20] to enhance classification performance and accuracy by discarding irrelevant attributes. Nevertheless, these studies did not consider the selected features' stability when applied in a situation with a different location and time heterogeneity. Moreover, for traffic identification at the network edge in real-time, a minimal number of features must be used to improve the classification throughput on edge devices such as middleboxes.

Thus, this work proposes a feature selection method for network traffic classification named Ensemble Weight Approach (EWA) for selecting robust statistical flow features for Internet traffic classification that are robust. The proposed feature selection method first generates candidate features using conventional feature selection methods, ranking each feature combination, and searching for the best features. Extensive simulations have been conducted using publicly-available traces from the University of Cambridge to evaluate the proposed EWA feature selection. EWA selects fewer features for machine learning classification of Internet traffic that are stable irrespective of the dataset's spatial and temporal differences, improving the overall traffic classification.

The remainder of this paper is organized as follows. Section II discusses similar feature selection methods, particularly for network traffic classification. Section III presents the proposed feature selection method. Section IV describes the experimental setup, while Section V discusses the results. Section VI concludes the paper and recommends future works.

II. RELATED WORK

This section discusses similar feature selection methods, particularly for network traffic classification. We also present a comprehensive review of state-of-the-art feature selection techniques for network traffic classification.

A. ML TRAFFIC CLASSIFICATION

One of the techniques that can be applied to IoT is ML. ML is a group of robust strategies for data mining and knowledge discovery [21], [22]. The first work using this technique was [23]. The conventional structure for creating ML models involves sampling the training dataset, extracting features, selecting informative features, and creating the generative model. Once the generative model has been generated, network traffic can be classified based on the preset classes defined during training.

Feature extraction is a method of extracting features that can distinguish a data class over the others. In the case of network traffic classification, distinct attributes such as port [4] and packet inter-arrival time and flow statistics [24] can be used as the classification features. However, the cardinality of possible features can be huge. While classifier training can be done offline, many features will result in a large generative model and require a big memory footprint. Furthermore, extracting a large number of features in real-time classification is not realistic. Hence, feature selection (FS) is required to boost both effectiveness and efficiency since it discards less informative or irrelevant features that benefit both the training and classification phases.

B. THE USE OF FEATURE SELECTION

In machine learning, FS is a commonly used technique in data preprocessing. FS methods aim to identify and choose a subset of features to describe the data concept effectively. Simultaneously, FS can reduce the effects of noise and unrelated attributes to yield a good prediction of data class [17], [25], [26]. Traffic identification can greatly benefit in terms of accuracy and other performance metrics by utilizing the most significant features [27]. The selection of relevant features for network traffic identification is non-trivial due to:

- It requires a good understanding of the traffic engineering domain to identify which features are relevant.
- Datasets may contain irrelevant and redundant features that considerably reduce classification accuracy.
- Efficiency of the identifiers decreases when selecting a huge number of attributes. The storage requirement is increased, and time taken for training and testing of the model is also increases [28].

Recently, FS strategies are extensively deployed in many applications, such as identifying informative genes [29], bioinformatics [30], and text categorization [31]. The objectives of the algorithms used for extracting features may differ. However, they all have many similarities [32]:

- To find the minimal size feature subgroup is fundamental and sufficient to the target concept [33].
- The ability to choose a subgroup of features from a large collection, in which the criterion value can be optimized over every subgroup [34].
- The right choice of subclass features to increase identification accuracy. Reducing the structure of chosen features and not tampering with the built model's prediction accuracy [35].
- Selecting a small group can result in class distribution given only values of the selected features, which can closely represent the original distribution [35].

Furthermore, FS process evaluation can be achieved with four basic stages: subset creation and assessment, termination criterion, and result validation [36]. The process starts with

subclass creation employing a particular search approach to yield candidate feature subsets. Subsequently, every candidate subgroup is examined using specific examination conditions and related to the previous best result. The obtained result becomes the best result if it outperforms the previous best. The procedure for subset creation and examination continues until the termination condition is fulfilled. Lastly chosen best feature subgroup is authenticated by previous information or test data. The search approach and assessment condition are two vital factors for the study of FS.

Subset creation starts with a search point, that could be an empty set, whole set, or a randomly created subset. In the beginning, it can lookup feature subgroups from random directions. In the forward search, features are inserted individually, whereas in the backward search, the least significant feature is detached based on the valuation criterion. Random search includes or removes features in random to evade being trapped into a local maximum.

C. FEATURE SELECTION MODELS

FS processes can be categorized into two main methods - filter and wrapper methods [37]–[39]. The filter approaches or feature ranking methods can use the wrapper approach to rank features too. A filter-based FS can return a subset of features, e.g., Correlation-based FS method (CFS). These techniques' attractive nature is centered on their simplicity, scalability, and good empirical success [14]. Feature ranking is effective because it involves only computation and sorting of scores. The subsets of the main features can be chosen based on feature ranks to create a classifier. Some filter techniques employ ranking conditions based on information-theoretic criteria including information gain (IG) [40], Gain-Ratio (GR) [27], mutual information [14], and entropy-based measure [41], whereas some use statistics, such as Chi-squared statistics [42], T-statistics [43], F-statistics [44], MIT correlation [45], and Fisher criterion [46].

The wrapper approaches [47], [48] rely on identifying informative features for obtaining a feature subset. Wrappers exploit the performance learning machine to appraise the value of feature subgroups. The wrapper FS techniques can produce high identification accuracy for a specific identifier at the expense of high computational complexity and less generalization of the selected features on other identifiers. The wrapper techniques commonly surpass filter technique with regards to the accuracy of the learning machine, which could be categorized as sequential selection algorithms ((SFS), sequential backward (SBFS), and sequential forward floating selection (SFFS)) and heuristic search algorithms (genetic algorithm [49]).

The other group of FS is hybrid methods. Every feature evaluation measure (EM) is equipped with distinct advantages and disadvantages. Some hybrid procedure FS techniques include filter and wrapper [39], [50]. Lately, the hybrid approach has been widely explored for FS due to its global optimization abilities [51]. The hybrid method proposed in [29] applied rank, which grouping to associate various FS approaches. These features were combined using a weighted sum from every component rankings acquired from a distinct FS mechanism. This shows that a combination scheme performs better than individual FS techniques. Moreover, Rogati and Yang [52] prove that the increase in performance was achieved by merging several feature selectors.

Moreover, all these methods can be represented in the space of features according to the evaluation measures (EM), generation of successor (GoS), and search organization standards. Generation of successor and Search organization are grouped as generation procedure. These three characteristics are described as follows.

- EM is a function used to evaluate the generated successor.
- GoS is a mechanism that proposes a successor of the current hypothesis. Different operators can be considered to generate a successor: Forward, Backward, Compound, Weighting, and Random.
- Search algorithm is used to drive the FS process using one of these strategies: sequential, exponential, or random strategy.

Moore et al. in [53] used the Fast Correlation Based Filter (FCBF) feature selection technique for feature reduction and Naive Bayes algorithm to measure the significance of the feature reduction. The overall classification accuracy result based on features subsets is 84.06%, obtained by using all features. Jun et.al., in [54] used two feature subsets to create a classified traffic. The work employed flow features subsets on Support Vector Machine (SVM). Training time was reported at 40 seconds, while the classifier accuracy is 70%. In [55] classified traffic using SVM and random search algorithm for features reduction. The proposed method did not use UDP traffic, even though network traffic is composed of TCP and UDP packets.

Zhang et al. [16] proposed WSU AUC and SRSF FS algorithms. WSU AUC was employed to select features from high dimensional imbalanced data. This work used ten Cambridge datasets, UNIBS, and CAIDA datasets and applied the C4.5 decision tree and NBK machine learning algorithm (batch learning method) to evaluate proposed FS algorithms. This method computes the value of WSU on each feature and the classes and removes redundant features depending on the specific three-shot. This method also used the SRSF method to select the robust features that depend on frequency weight. This work selected three server port features, the total number of bytes sent in the initial window and minimum segment size observed, hence achieved an accuracy of more than 94%.

D. CHALLENGES IN FEATURE SELECTION FOR TRAFFIC CLASSIFICATION

The key challenge for selecting features is preserving the appropriate features subset for accurate traffic identification. Traffic classification accuracy is associated with a small number of appropriate features [13], [15]–[17]. Various FS

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037492, IEEE Access

methods select various sets of significant features, but they do not always select the same number of significant features. These are challenging due to:

- Representative influence of a specific FS approach may limit its search space, which hinders achieving an optimal subset.
- Various FS approaches may produce feature subgroups that can be termed as local optimal in the space of feature subsets.
- A collective method can give an improved approximation to optimal subset or ranking of features, which is not frequently applicable with a single feature selection technique.

Moreover, a broad analysis is required to provide information or knowledge for the main factors affecting the robustness of the FS procedures. Al Harthi et al. [56] proposed an approach named global optimization algorithm (GOA) it was focused on the stability issues. This approach depends only on the frequency of the selected feature (ignore the robustness of the selected feature) and consider Round-Trip Time (RTT) features as part of selected subset features, which depend on location [57].Nevertheless, it would be ideal to ensure the robustness of feature subset (accurate regardless of location and time heterogeneity and selection of a small relevant number of features). This is important to build traffic identification.

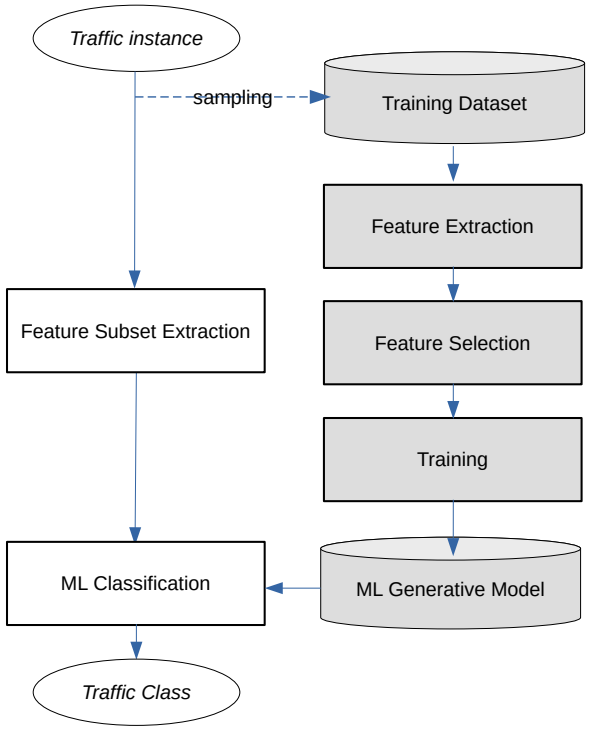

FIGURE 1. Generic stages for ML traffic classification. The shaded tasks are for training the ML generative model that can be done offline.

III. PROPOSED ENSEMBLES WEIGHT AVERAGE (EWA) FEATURE SELECTION

The conceptual illustration of traffic classification is shown in Figure 1. This framework comprises the learning model that learns from the sampled datasets and the classifier model that classifies incoming traffic based on the learned classifier model. A traffic instance (packet or flow, depending on stateless or stateful processing) is represented by several features that can measure varying aspects of such an instance. A flow refers to a group of packets sharing same 5-tuples (source and destination IP, source and destination port, and transport protocol). Flow can be represented by UDP or TCP packets.

Generally, datasets (can be in the pcap format) are used as the classifier's training sample. Then, the FS selects the relevant feature subsets to the target protocol or application (in this case, network traffic classification). The learning model is then learned based on the selected feature subsets of all training instances.

As previously mentioned, the hybrid method combines features based on a weighted sum from every component rankings acquired from a distinct FS mechanism. This approach is shown to perform better than individual FS techniques. The EWA method consists of three main stages: Evaluation of individual FS methods and feature pool generation, weighted ranking of features, and searching an optimal features subset, as shown in Figure 2. The first stage involves feature extraction and the formation of a feature pool from outputs of individual FS methods (wrapper and filter FS methods). The cutpoint of twenty features is used as the stopping criterion. The cutpoint value can be changed accordingly. Since the EWA aims to select the fewest possible traffic classification features, the cutpoint is set to twenty. In the second stage, the selected features are ranked, and features observed in different datasets will be given higher ranks. In the third stage, EWA applies one widely used sequential search strategy (SFS) (Sequential Forward Selection (SFS)) [58] to remove irrelevant and redundant features from the initial selected features pool.

STAGE 1: FEATURE POOL GENERATION

This stage evaluates the stability of each feature subset generated by the respective FS technique. Each FS technique generates non-unique feature subsets when applied to the different training datasets. Note that a distinct FS technique uses a distinct method to create feature subsets. The selected features are then evaluated using an ML classifier, in this work the naive Bayes classifier is appiled to evaluate the accuracy of each dataset. Selecting optimal features across the different locality and time heterogeneity is difficult. Hence, to make the best of the various FS methods, EWA uses multi-feature selection methods on multiple datasets to create the initial pool of multiple feature subsets. Accuracy and Stability are used as the criterion to select the candidate FS methods. These selected feature subsets are used to create the initial features pool. Unselected features by any of the FS techniques are removed.

Assume a set of training datasets, $D = \{D_1, D_2, ..., D_{|D|}\}$, k is the number of candidate FS methods $FS =$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037492, IEEE Access

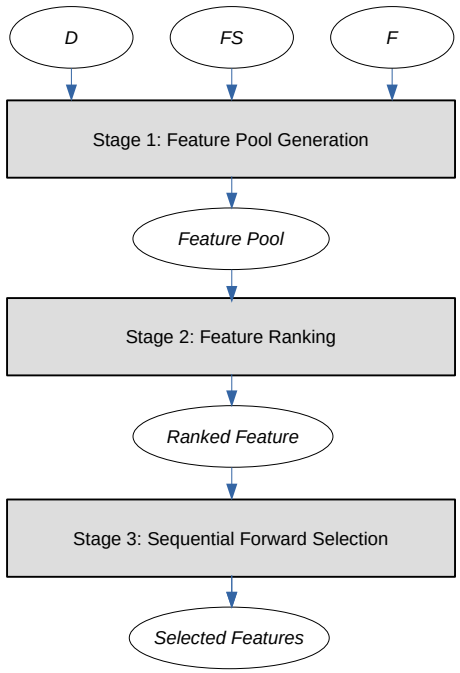

FIGURE 2. Generic stages for ML traffic classification. The shaded tasks are for training the ML generative model that can be done offline.

$\{FS_1, FS_2, ..., FS_k\}$, and $F = \{f_1, f_2, ..., f_{|F|}\}$ is the potential features that can be used for traffic classification. Moore et al. proposed 248 porential features that be used for network traffic classification [19]. Let $P_{pool} = \emptyset$ be the initial features pool and P_k is the the X best ranking features for FS_k, where X is features cutpoint. Each P_k is evaluated using a ML classifier, in this work the naive Nayes classifier, to evaluate its accuracy Ac_k and Stability St_k. Algorithm 1 shows the initial features pool.

Algorithm 1 Feature Pool Generation

```
Datasets D = {D_1, D_2, ..., D_|D|}
FS methods FS = {FS_1, FS_2, ..., FS_k}
Feature pool P_|pool = ∅
for i = 1 to |D| do
    for j = 1 to k do
        Generate P_k
        Evaluate Ac_k and St_k
        Select first X features in ranking
        P_pool ← P_pool ∪ P_k
    end for
end for
return P as the candidate features
```

Using the cross-validation, the Accuracy $Ac_{k,i}$ due to the selected features by FS_k on dataset D_i is given as

$$Ac_{k,i} = \frac{t_p + t_n}{t_p + t_n + f_p + f_n} \quad (1)$$

$$Ac_k = \text{avg}(Acc_{k,i}) \quad \forall D_i \in D \quad (2)$$

where t_p, t_n, f_p, and f_n respectively represents true positive, true ngative, false positive, and false negative. Accuracy $Ac = [0, 1]$, where $Ac_k \rightarrow 1$ shows accurate traffic classification whereas $Ac_k \rightarrow 0$ indicates inaccurate traffic classification.

Stability St is a measurement to indicate the robustness of the selected features regardless of traffic data variations. A certain FS method may generate different feature sets on datasets collected in different periods or locality due to concept drift. Therefore, it is critical to select features that can yield high prediction Accuracy and better relative Stability over different samples. This study employed the stability measure suggested by [20] to evaluate the distinct feature selection methods.

A FS may respectively generate P_a and P_b feature subsets from datasets D_a and D_b, where both maybe unidentical. Let $P_k = P_a \cup P_b$. The stability St_k of the selected features by FS_k over the two datasets can be estimated according to [20] as:

$$St_k = [1 - RU(P_k)] \times 100 \quad (3)$$

$$RU(P_k) = \frac{H(P_k)}{log(|F|)} \quad (4)$$

$$H(P_k) = \frac{1}{|F|} \sum_{i=1}^{|F|} -\frac{n_j^i}{k|D_j|} log\left(\frac{n_j^i}{k|D_j|}\right) \quad (5)$$

where $|F|$ is the total number of features, n_j^i is the frequency of specific feature f_i observed across different datasets D_j.

STAGE 2: WEIGHTED RANKING OF FEATURES

EWA is based on a weighted ranking measure to select robust features using multiple individual FS methods on different traffic datasets. The idea behind this as a class is superiority over that of individual FS methods, where the most significant features for network traffic classification are probably be endorsed by most FS methods.

A weighted ranking measure for each feature f_i is R_{f_i}, which is the likelihood that f_i is selected by multiple FS methods in different traffic datasets (or none at all), as shown in Equations (2). The mean value of R_{f_i} in Equation (5) shows high optimality when $\text{avg}(R_{f_i}) \rightarrow 1$, whereas $\text{avg}(R_{f_i}) \rightarrow 0$ indicates low optimality.

Let $|D|$ denotes the cardinality of traffic datasets D, where k represents the total number of FS methods used on a single dataset. A weighted ranking for each feature f_i is given as:

$$R_{f_i} = \frac{1}{k|D|} \sum_{j=1}^{|D|} \sum_{z=1}^{k} O_{i,j,z} \quad (6)$$

$$O_{i,j,z} = \frac{X - L_{i,j,z}}{X} \quad (7)$$

$$\text{avg}(R_{f_i}) = \frac{\sum_{i=1}^{|F|} R_{f_i}}{|F|} \quad (8)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037492, IEEE Access

where $O_{i,j,z}$ denotes the weight of feature f_i dependent on its location $L_{i,j,z}$ w.r.t. cutpoint value X for each D_j and FS_z. The lower the value of $L_{i,j,z}$ indicates its high significance.

An optimal threshold value is needed for selecting features that are stable and have high weighted ranks, which are sufficiently unique and reliable. As an example, a feature with a high average ranking weight is considered sufficiently reliable. The threshold $B = R_{f_i} - \text{avg}(R_{f_i})$ is determined through experimentation. The higher value of B may not necessarily result in higher Accuracy as too few features may be used to classify network traffic.

In the second algorithm, firstly, the average weight measures of the features f_i are computed. Then each feature which has average weight measures more than or equal threshold is selected and kept into the set of the best stable features subset P_{ranked} subset. Finally, important features containing indispensable information about the original features are selected.

Algorithm 2 Weighted Ranking of Features

```
Features set F = {f_1, f_2, ..., f_|F|}
Features pool P_pool (stage 1)
Ranked features set P_ranked = ∅
Threshold B
for f_i ∈ P_pool do
    Compute B_i = R_{f_i} − avg(R_{f_i})
    if B_i ≥ B then
        P_ranked ← P_ranked ∪ f_i;
        Update B = B_i
    end if
end for
return P_ranked
```

STAGE 3: SEARCH THE BEST FEATURES SUBSET USING SEQUENTIAL FORWARD SELECTION

In this stage, we apply the wrapper approach to identify the best candidate features as a good search technique. The techniques, in general, are classified into three groups: randomized, exponential, and sequential. This research considers a widely used Sequential Forward Selection (SFS), a sequential search strategy [58]. SFS selects the best combination of subset features for extraction. The selection process begins with an empty set and continuously adds a single feature from the superset to the subset when the Accuracy increases.

Table 1 illustrate the modified SFS to create a selected features from a ranked features subest. In this case, $\{f_1, f_2, f_3, f_4\}$ are selected as the features to be used in network trafffic classification.

IV. EXPERIMENTAL SETUP

This section describes the validation of EWA compared to other feature selection methods.

TABLE 1. Procedure of SFS

Steps	Feature Set	Ac	Selected features
Step 0	..., ..., ..., ...	0	$\emptyset$
Step 1	f_1, ..., ..., ...	30	f_3
	..., f_2, ..., ...	25	
	..., ..., f_3, ...	35	
	..., ..., ..., f_4	28	
Step 2	f_1, ..., f_3, ...	30	f_2
	..., f_2, f_3, ...	45	
	..., ..., f_3, f_4	40	
Step 3	f_1, f_2, f_3, ...	65	f_4
	..., f_2, f_3, f_4	70	
Step 4	f_1, f_2, f_3, f_4	75	f_1

A. VALIDATION PROCEDURE

The validation procedure involves evaluating the proposed EWA feature selection compared to the IG [59], FCBF [53], and GOA method [56] in term of Accuracy (Ac), Stability (St), and Root Mean Squared Error (RSME).

The following software and tools were used to achieve the set objectives of this work:

- Batch learning algorithms are frameworks that facilitate the selection of the appropriate attributes for the identification of Internet traffic. Naive Bayes (NB) was used as classifiers. These classifiers have been successfully employed in various works tackling traffic classification [60]. They were executed in Weka open-source platform [**?**].
- Weka [61] a data mining software was used to implement the selection of select suitable and correct traffic features.
- A laptop with Intel Core i7-5500U processor, 8 GB RAM, and 1 TB HDD was used for validation purposes.

B. DATASET

EWA was evaluated using the widely acceptable traffic datasets from the University of Cambridge [19] (dataset D_1 to D_10). This dataset is among the largest network traffic traces, which is publicly-available and assembled by a high-performance network monitor over different periods from two different network sites. The sites are designated as Site A and Site B, with each site hosts about 1,000 Internet-connected users through a full-duplex Gigabyte Ethernet link. A high-performance network monitors the full-duplex traffic for each traffic set on this connection. Table 2 summarizes the datasets. For the implementation, we used the Weka data mining tool [61]. In the Cambridge dataset case, the early stage-packet statistic is not available without access to all raw packets. Hence, the complete flow statistics are used. To give an impartial assessment of all datasets, the Cambridge dataset's mean attributes were recomputed to obtain the total attributes.

C. EVALUATION METRICS

Primarily, the proposed EWA is evaluated in terms of Accuracy and Stability as described in Section III. To measure

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037492, IEEE Access

TABLE 2. Cambridge Dataset [19]

Datasets	#flows	Size (MB)
D_1	24,863	29.7
D_2	23,801	28.3
D_3	22,932	27.5
D_4	22,285	26.6
D_5	21,648	25.8
D_6	19,384	23.1
D_7	55,835	66.0
D_8	55,494	65.6
D_9	66,248	78.3
D_{10}	65,036	77.1

EWA in relation to other similar works, Similarity Si and Root Mean Square Error (RSME) are also used.

This research evaluates the similarity with respect of the accuracy of FS techniques when classifying traffic datasets collected from different time and location. The similarity (Si) in term of accuracy between two candidate datasets, D_a and D_b is defined as

$$Si(D_a, D_b) = 1 - \frac{1}{2}\sum_{c=1}^{|C|} |Ac_{a,c} - Ac_{b,c}| \tag{9}$$

where C is the set of ML classifiers used to evaluate the datasets. The Similarity measure takes values $Si = [0, 1]$, where the value close to zero indicates low similarity in accuracy across multiple datasets and ML classifiers, while the value approaching 1 indicates high similarity in terms of accuracy.

Root Mean Square Error (RMSE) is a quadratic scoring to measure the average magnitude of error. RMSE gives a relatively high weight to large errors. Hence, RMSE is most useful when large errors are particularly undesirable.

V. RESULTS AND DISCUSSION

Based on EWA stages that have been described in Section III, this section explains the results from EWA:

- We evaluated seven FS algorithms in order to choose the best top methods.
- The selected FS methods are applied to generate the features pool.
- The weight of the features in the features pool is used to select the best features depending on the threshold.
- Sequential Forward Selection method and Naive-Bayes classifier are applied to select the best combination subset of features.
- Lastly, we compare EWA method with FS methods: IG [59], FCBF [53] and GOA method proposed in [56].

A. STAGE 1 RESULT

We evaluated seven FS techniques i.e., GainRatio (GR), Chi-square (Chi), information gain (IG), Correlation Attribute Eval (CAE), CV Attribute Eval (CV AE), Principal components (PC), and Consistency Subset Eval (CSE) on four Cambridge dataset which are D_1, D_3, D_6 and D_{10}. The cutpoint $X = 20$ is applied for the ranking method. After

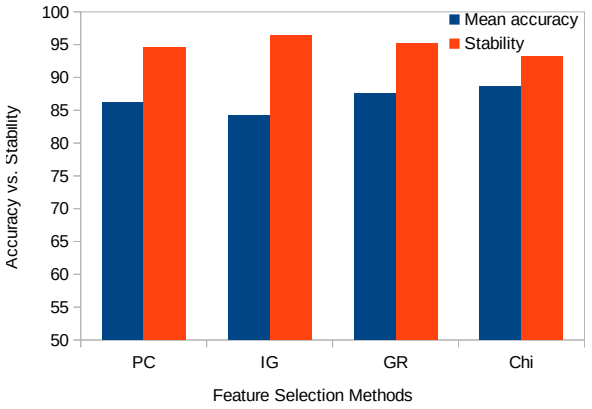

FIGURE 3. Accuracy and Stability of multiple existing FS methods.

that, FS methods that achieved higher mean accuracy were selected.

Table 3 presents a comparison of classification accuracy for seven (7) FS methods on four Cambridge datasets (D_1, D_3, D_6 and D_{10}). Hence, an FS strategy with high mean accuracy is preferred. It is worthy of note the FS methods that give higher accuracy are ranked as follows: Chi-square, PC, GR, IG, CSE, CAE, and CV AE; as presented in Table 3. As a result, we select Chi-square, PC, GR, and IG methods for Stage 1 of EWA.

TABLE 3. The average accuracy $\mathrm{avg}(Ac)$ for conventional FS methods.

Dataset	IG	GR	Chi	CAE	CVAE	PC	CSE
D_1	60.34	80.81	80.80	73.04	76.12	80.25	77.35
D_3	93.96	84.11	82.39	84.81	29.46	85.08	86.00
D_6	96.05	91.11	93.08	91.95	90.91	90.02	86.56
D_{10}	85.44	89.7	91.80	84.90	80.12	89.25	77.35
$\mathrm{avg}(Ac)$	84.24	87.52	88.61	84.08	72.79	86.15	81.81

The selected FS techniques (Chi-square, PC, GR, and IG methods) are compared based on their accuracy and stability (see Figure 3). None of the FS methods outperformed the others in most cases as there is no available FS technique that can satisfy both criteria (stability and accuracy). For instance, the performance of Chi-square FS was good on the accuracy metric but poor on the stability metric. Meanwhile, PC was poor on both metrics, while IG performed well on stability but poor on accuracy.

Therefore, it is concluded that each of the evaluated FS methods has its advantages and disadvantages when measured in terms of accuracy and stability. Our motivation for proposing a ranking method based on multi-criterion methods is to identify a stable and optimal subset of features that help traffic classifiers perform well across different times and locations. In this stage also, we evaluated 248 features (see [19]) using four FS techniques (GR, Chi-square, IG, and PC). The experiment utilized ten Cambridge dataset D_1 to D_{10} with cutpoint equals twenty is applied, and the best 20 features in the ranking are selected.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037492, IEEE Access

B. STAGE 2 RESULT

In this stage, we compute the mean ranking weight of all 248 features $F = \{f_1, f_2, ..., f_{248}\}$ by using four FS techniques on the ten Cambridge datasets and filter out all features that have average weight $R_{f_i} \leq 0.005$, which reduces the number of features from 248 to 32 features as tabulated in Table 4. Therefore, features with higher mean weight are desired. The features f_1, f_{95}, f_{96}, f_{180} and f_{187} achieve a higher mean weights.

Table 4 shows the threshold value *(B)* for all selected 32 features. Here we set *(B)* to select the best features depending on the best result during evaluation using the Naive Bayes classifier and the Cambridge datasets. Table 5 tabulates threshold B values and the number of features and their respective accuracy for each range of *B*. The results explain the value of $B \geq 0.054$ is the best accuracy than other values of *B*.

TABLE 4. Weighting matrix result with $\text{avg}(R_{f_i}) = 0.196$

f_i	R_{f_i}	B	f_i	R_{f_i}	B
f_1	0.950	0.754	f_{95}	0.486	0.290
f_{24}	0.230	0.230	f_{96}	0.415	0.219
f_{43}	0.061	-0.134	f_{101}	0.116	-0.080
f_{47}	0.065	-0.131	f_{125}	0.195	-0.001
f_{82}	0.172	-0.023	f_{133}	0.204	0.008
f_{83}	0.250	0.054	f_{135}	0.082	-0.114
f_{84}	0.147	-0.049	f_{136}	0.055	-0.141
f_{86}	0.142	-0.054	f_{137}	0.168	-0.028
f_{90}	0.118	-0.078	f_{143}	0.106	-0.090
f_{93}	0.179	-0.017	f_{145}	0.086	-0.110
f_{94}	0.136	-0.060	f_{147}	0.108	-0.088
f_{149}	0.105	-0.091	f_{179}	0.108	-0.088
f_{151}	0.136	-0.060	f_{180}	0.452	0.252
f_{159}	0.062	-0.062	f_{184}	0.279	0.083
f_{167}	0.181	-0.015	f_{186}	0.260	0.064
f_{177}	0.207	0.012	f_{187}	0.340	0.144

TABLE 5. The threshold values and their relationship with the number of features and accuracy.

Threshold range	# features	avg(Ac)
$B \geq -0.141$	32	78.540
$B \geq -0.091$	26	82.800
$B \geq -0.028$	17	86.080
$B \geq 0.008$	10	89.790
$B \geq 0.054$	8	90.830

C. STAGE 3 RESULT

In this stage, the Sequential Forward Selection (SFS) method and the Naive Bayes classifier are applied to select the best feature combinations as the classification features. The SFS method begins with an empty set and continuously adding a single feature at any time until all possible combinations are tested. Table 6 explains the selection of these features.

D. PERFORMANCE COMPARISON WITH OTHER FS METHODS

Not to be biased with the proposed metrics, EWA is compared with full features, baseline FS methods: IG [59], FCBF [53] and GOA method proposed in [56]. The proposed method was tested and validated using the same metrics.

TABLE 6. Selected features based on features described in [19].

Feature	Feature name	Feature description
f_1	server_port	Port Number at server
f_{95}	initial_window_bytes_ab	The total number of bytes send in the initial window (client to server)
f_{96}	initial_window_bytes_ba	The total number of bytes send in the initial window (server to client)
f_{180}	var_data_wire_ba	Variance of bytes in (Ethernet) packets
f_{187}	var_data_ip_ba	Variance of total bytes in IP packet

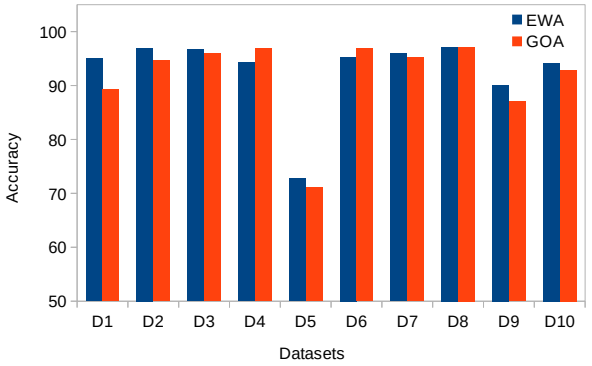

FIGURE 4. Accuracy of EWA compared to GOA.

TABLE 7. Accuracy (%) for different FS methods for different dataset2s.

Datasets	FCBF	IG	GOA	EWA	FF
D_1	92.06	60.34	89.22	95.05	57.89
D_2	50.44	87.30	94.62	96.95	53.90
D_3	93.65	93.96	95.97	96.63	84.45
D_4	95.40	81.92	96.89	94.31	74.51
D_5	64.22	72.52	71.06	72.71	63.12
D_6	87.29	96.05	96.90	95.27	90.07
D_7	94.62	62.47	95.15	95.97	51.86
D_8	92.06	71.15	97.15	97.16	58.35
D_9	44.19	63.04	87.02	90.11	67.44
D_{10}	93.83	85.44	92.86	94.18	55.44
avg(Ac)	80.77	77.41	91.46	92.83	65.70

Table 7 presents results of comparison between the proposed method and full features (FF), baseline FS methods: IG [59], Fast Correlation-Based Filter (FCBF) [53] and GOA method proposed in [56]. EWA improves mean accuracy up to 4.2% using Naive Bayes for the 10 Cambridge dataset, and at the same time, it uses the smallest number of features (5 features) compared with others. Figure 4 shows EWA's accuracy achieves a slight improvement over the GOA method, while full features perform poorly.

Table 8 shows the comparison in terms of RMSE between EWA and GOA. The results indicate that the EWA approach achieved slight improvement overall compared to other FS methods for the ten Cambridge datasets, as shown in Table 8, while full features perform poorly. For the RMSE comparison between EWA and GOA, the EWA approach has achieved slight improvement over the GOA method, as

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037492, IEEE Access

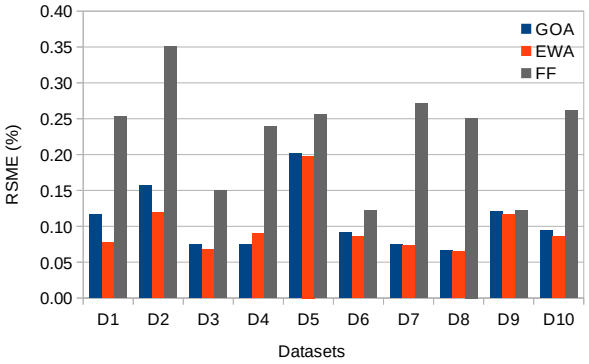

FIGURE 5. RMSE comparison between EWA, GOA, and full-features.

shown in Figure 5.

TABLE 8. Root mean squared error (RMSE) comparison between EWA, full-feature, and other FS methods (in %).

Datasets	FCBF	IG	GOA	EWA	FF
D_1	0.097	0.222	0.116	0.078	0.253
D_2	0.232	0.122	0.157	0.119	0.351
D_3	0.090	0.093	0.075	0.068	0.150
D_4	0.082	0.153	0.075	0.090	0.239
D_5	0.214	0.204	0.201	0.198	0.256
D_6	0.125	0.077	0.091	0.086	0.122
D_7	0.087	0.200	0.074	0.073	0.271
D_8	0.093	0.186	0.066	0.065	0.251
D_9	0.221	0.217	0.121	0.116	0.122
D_{10}	0.086	0.143	0.094	0.086	0.261
Avg(RSME)	0.182	0.161	0.106	0.090	0.228

Figure 6 shows the comparison between EWA, full features (FF), and other FS techniques (IG, FCBF, and GOA) in accuracy and stability. The full-feature performs very well on the stability but fares poorly in accuracy. The full features set contains many redundant and irrelevant features. Other FS methods such as IG performed poorly on accuracy but performed equally well on stability, while FCBF performs poorly on both metrics. GOA and EWA outperform the other FS techniques (i.e., IG and FCBF) on both stability and accuracy metrics, as various FS techniques are incorporated in GOA and EWA to select different groups of relevant features.

Conventional FS methods may not agree on the same relevant features for these reasons: Different FS methods may select feature subsets that can be considered local optimal in the feature subsets space.

- The search space of any FS technique may be restricted by the technique's representative power such that it may be impossible to reach the optimal subset.
- The combination of more than one approach can produce a better ranking of features or a better approximation to the optimal subset.
- In most cases, EWA outperforms all methods in terms of stability and accuracy. Although GOA and EWA have similar stability, EWA outperforms GOA because EWA is based on a weighted ranking measure that allows the selection of robust features from multiple FS techniques on different traffic datasets.

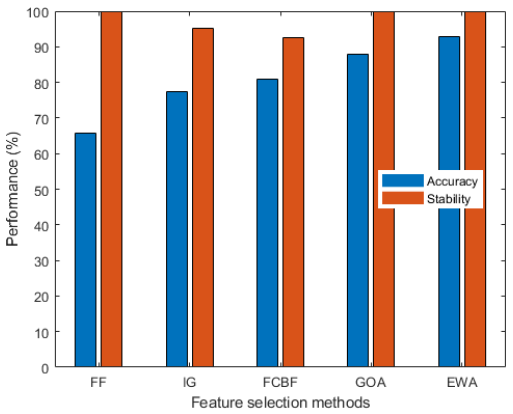

FIGURE 6. Accuracy and Stability of EWA compared to GOA and other existing FS methods.

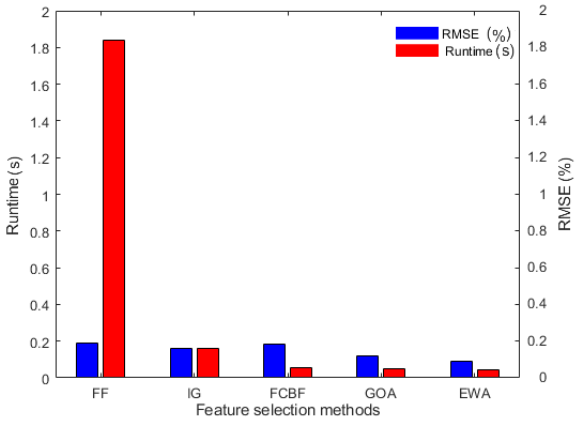

FIGURE 7. RMSE versus Runtime of EWA compared to GOA and other existing FS methods.

Figure 7 shows the comparison of full features (FF), FS techniques (IG [59], FCBF [53]) and GOA methods [56] compared to the proposed EWA method in terms of RMSE and time to build the model (runtime (in seconds)). As a result, full features have very high RMSE and Runtime (i.e., using full features). FCBF generates high RMSE and low Runtime, while IG performs equally poorly on both. GOA and EWA methods outperform full features and selected FS techniques (IG, FCBF) in RMSE and Runtime criteria. In most cases, EWA performs better than GOA in terms of RMSE as GOA depends only on the selected feature's frequency. Both EWA and GOA have similar Runtime (s) due to both methods selected only five features for classification.

Tables 9 and 10 show the comparison of EWA with GOA in terms of similarity and accuracy. Naive Bayes and decision tree J48 ML classifiers are applied on Cambridge datasets D_1 and D_2, collected at different times (Table 9) and datasets D_1 and D_2 collected from different locations (Table 10). Results

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037492, IEEE Access

show that EWA performs better than GOA in similarity and accuracy as EWA is based on a weighted ranking measure. This allows a selection of features selected by multiple FS techniques from different traffic datasets with different time and location heterogeneity.

TABLE 9. Similarity in accuracy for dataset collected at different times.

	GOA		EWA	
	J48	NB	J48	NB
D_1	0.996	0.842	0.997	0.955
D_2	0.998	0.946	0.998	0.970
avg(Ac)	0.997	0.894	0.998	0.962
Si	0.949		0.983	

TABLE 10. Similarity in accuracy for dataset collected from different locations.

	GOA		EWA	
	J48	NB	J48	NB
D_7	0.998	0.952	0.998	0.960
D_{10}	0.990	0.929	0.997	0.941
avg(Ac)	0.997	0.940	0.998	0.951
Si	0.971		0.983	

The simulation results indicate that EWA can perform the selection of stable features that can be applied at different times and location heterogeneity. However, in some practical traffic classification use-cases that require modularity and scalability, such as in hierarchical classification [62], time and location heterogeneity are undesirable. EWA can still be used as the feature subsets are dependent on the used datasets. By categorizing training datasets, different feature subsets for hierarchical traffic classification can be obtained.

VI. CONCLUSION

This paper contributes to the selection of robust feature subsets for the identification of Internet traffic. The Ensemble Weighted Approach (EWA) feature selection method was proposed to select robust subset features for Internet traffic identification. The results of the experiments proved that no singular feature selection technique could perform well on all datasets. Based on this fact, we suggested a method that relies on the positives of the individual FS methods to obtain a robust method. The simulation results on real datasets illustrate EWA's capability to identify robust subset features for Internet traffic identification. Our findings also show that EWA improves mean accuracy up to 1.3% and, at the same time, reduced RMSE up to 0.016 uses a smaller number of features that directly contribute to improving Runtime up to 0.003 seconds). Selected features can build stable traffic identification models that remain accurate regardless of location and time heterogeneity with high similarity above 97%.

For future works, we plan to further analyze EWA for the early estimation of statistical flow features. This is important for real-time traffic identification as only certain features can be extracted on the wire with the limited flow or packet observability. We also plan to enhance ML classification with incremental learning, as there is a need to propose forgetting to enhance traffic classification accuracy over time by removing uninformative features when concept drift happens. Also, a real-time traffic detection system can be integrated with any network traffic management.

REFERENCES

[1] H. Jiang, A. W. Moore, Z. Ge, S. Jin, and J. Wang, "Lightweight application classification for network management," in Proceedings of the 2007 SIGCOMM Workshop on Internet Network Management, 2007, pp. 299–304.

[2] D. Moore, K. Keys, R. Koga, E. Lagache, and K. C. Claffy, "The Coralreef software suite as a tool for system and network administrators," in Proceedings of the 15th USENIX conference on System administration, 2001, pp. 133 –144.

[3] A. W. Moore and K. Papagiannaki, "Toward the accurate identification of network applications," in International Workshop on Passive and Active Network Measurement, 2005, pp. 41–54.

[4] V. Paxson, "Bro: A system for detecting network intruders in real-time," Computer Networks, vol. 31, no. 23, pp. 2435 –2463, 1999.

[5] S. Sen, O. Spatscheck, and D. Wang, "Accurate, scalable in-network identification of p2p traffic using application signatures," in Proceedings of the 13th International Conference on World Wide Web, 2004, pp. 512–521.

[6] X.-B. Liu, J.-H. Yang, G. Xie, and Y. Hu, "Automated mining of packet signatures for traffic identification at application layer with apriori algorithm," Journal of Communication, vol. 29, no. 12, pp. 51–59, 2009.

[7] D. Ariu and G. Giacinto, "A modular architecture for the analysis of http payloads based on multiple classifiers," in International Workshop on Multiple Classifier Systems, 2011, pp. 330–339.

[8] Z. Guo and Z. Qiu, "Identification peer-to-peer traffic for high speed networks using packet sampling and application signatures," in Proceedings of the 2008 9th International Conference on Signal Processing (ICSP 2008), 2008, pp. 2013–2019.

[9] C. Yin, S. Li, and Q. Li, "Network traffic classification via HMM under the guidance of syntactic structure," Computer Networks, vol. 56, no. 6, pp. 1814–1825, 2012.

[10] J. Zhang, Y. Xiang, Y. Wang, W. Zhou, Y. Xiang, and Y. Guan, "Network traffic classification using correlation information," IEEE Transactions on Parallel and Distributed Systems, vol. 24, no. 1, pp. 104–117, 2013.

[11] H. R. Loo and M. N. Marsono, "Online network traffic classification with incremental learning," Evolving Systems, vol. 7, no. 2, pp. 129–143, 2016.

[12] T.-S. Chou, K. K. Yen, and J. Luo, "Network intrusion detection design using feature selection of soft computing paradigms," International Journal of Computational Intelligence, vol. 4, no. 3, pp. 196–208, 2008.

[13] B. M. A. Abdalla, H. A. Jamil, M. Hamdan, J. S. Bassi, I. Ismail, and M. N. Marsono, "Multi-stage feature selection for on-line flow peer-to-peer traffic identification," in Asian Simulation Conference, 2017, pp. 509–523.

[14] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning Research, vol. 3, pp. 1157–1182, 2003.

[15] T. Auld, A. W. Moore, and S. F. Gull, "Bayesian neural networks for internet traffic classification," IEEE Transactions on Neural Networks, vol. 18, no. 1, pp. 223–239, 2007.

[16] H. Zhang, G. Lu, M. T. Qassrawi, Y. Zhang, and X. Yu, "Feature selection for optimizing traffic classification," Computer Communications, vol. 35, no. 12, pp. 1457–1471, 2012.

[17] H. A. Jamil, A. Mohammed, A. Hamza, S. M. Nor, and M. N. Marsono, "Selection of on-line features for peer-to-peer network traffic classification," in Recent Advances in Intelligent Informatics, 2014, pp. 379–390.

[18] R. Yuan, Z. Li, X. Guan, and L. Xu, "An SVM-based machine learning method for accurate internet traffic classification," Information Systems Frontiers, vol. 12, no. 2, pp. 149–156, 2010.

[19] A. Moore, D. Zuev, and M. Crogan, "Discriminators for use in flow-based classification," Tech. Rep., 2005.

[20] F. A. AlHarthi, Z. Tari, I. Khalil, A. Almalawi, and A. Y. Zomaya, "An optimal and stable feature selection approach for traffic classification based on multi-criterion fusion," Future Generation Computer Systems, vol. 36, pp. 156–169, 2014.

[21] T. T. Nguyen and G. Armitage, "A survey of techniques for internet traffic classification using machine learning," IEEE Communications Surveys & Tutorials, vol. 10, no. 4, pp. 56–76, 2008.

[22] H. A. Jamil, B. M. Ali, M. Hamdan, and A. E. Osman, "Online p2p internet traffic classification and mitigation based on Snort and ML," European Journal of Engineering Research and Science, vol. 4, no. 10, pp. 131–137, 2019.
[23] M. G. Schultz, E. Eskin, F. Zadok, and S. J. Stolfo, "Data mining methods for detection of new malicious executables," in Proceedings of the 2001 IEEE Symposium on Security and Privacy (S&P 2001), 2001, pp. 38–49.
[24] A. Monemi, R. Zarei, and M. N. Marsono, "Online NetFPGA decision tree statistical traffic classifier," Computer Communications, vol. 36, no. 12, pp. 1329–1340, 2013.
[25] X.-Y. Liu, Y. Liang, S. Wang, Z.-Y. Yang, and H.-S. Ye, "A hybrid genetic algorithm with wrapper-embedded approaches for feature selection," IEEE Access, vol. 6, pp. 22 863–22 874, 2018.
[26] S. Khan, A. Gani, A. W. Abdul Wahab, and P. K. Singh, "Feature selection of denial-of-service attacks using entropy and granular computing," Arabian Journal for Science and Engineering, vol. 43, no. 2, pp. 499–508, 2018.
[27] O. Henchiri and N. Japkowicz, "A feature selection and evaluation scheme for computer virus detection," in Data Mining, 2006. ICDM'06. Sixth International Conference on, 2006, pp. 891–895.
[28] A. L. Blum and P. Langley, "Selection of relevant features and examples in machine learning," Artificial intelligence, vol. 97, no. 1-2, pp. 245–271, 1997.
[29] H.-Y. Chuang, H. Liu, S. Brown, C. McMunn-Coffran, C.-Y. Kao, and D. F. Hsu, "Identifying significant genes from microarray data," in Proceedings of the Fourth IEEE Symposium on Bioinformatics and Bioengineering (BIBE 2004), 2004, pp. 358–365.
[30] E. P. Xing, M. I. Jordan, and R. M. Karp, "Feature selection for high-dimensional genomic microarray data," in Proceedings of the Eighteenth International Conference on Machine Learning, 2001, pp. 601–608.
[31] P. C. G. Fung, F. Morstatter, and H. Liu, "Feature selection strategy in text classification," in Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2011, pp. 26–37.
[32] M. Dash and H. Liu, "Feature selection for classification," Intelligent data analysis, vol. 1, no. 3, pp. 131–156, 1997.
[33] K. Kira and L. A. Rendell, "The feature selection problem: Traditional methods and a new algorithm," in Proceedings of the Tenth National Conference on Artificial intelligence (AAAI), vol. 2, 1992, pp. 129–134.
[34] P. M. Narendra and K. Fukunaga, "A branch and bound algorithm for feature subset selection," IEEE Transactions on Computers, vol. 9, no. C-26, pp. 917–922, 1977.
[35] D. Koller and M. Sahami, "Toward optimal feature selection," Stanford InfoLab, Tech. Rep., 1996.
[36] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 4, pp. 491–502, 2005.
[37] B. Ali, H. Jamil, I. Ismail, and S. Mohd Noor, "Hybrid filter/wrapper feature selection methods for detecting new malware variants," in Proceeding of International Science Postgraduate Conference Malaysia, 2015.
[38] R. Kohavi and G. H. John, "Wrappers for feature subset selection," Artificial Intelligence, vol. 97, no. 1-2, pp. 273–324, 1997.
[39] F. Provost and V. Kolluri, "Scaling up inductive algorithms: an overview," in Proceedings of the Third International Conference on Knowledge Discovery and Data Mining, 1997, pp. 239–242.
[40] Y. Liu, "A comparative study on feature selection methods for drug discovery," Journal of Chemical Information and Computer Sciences, vol. 44, no. 5, pp. 1823–1828, 2004.
[41] M. Dash, K. Choi, P. Scheuermann, Liu, and Huan, "Feature selection for clustering-a filter solution," in Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002), 2002, pp. 115–122.
[42] H. Liu and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes," in Proceedings of the Seventh International Conference on Tools with artificial intelligence, 1995, pp. 388–391.
[43] H. Liu, J. Li, and L. Wong, "A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns," Genome Informatics, vol. 13, pp. 51–60, 2002.
[44] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 27, no. 8, pp. 1226–1238, 2005.
[45] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri et al., "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," science, vol. 286, no. 5439, pp. 531–537, 1999.
[46] T. S. Furey, N. Cristianini, N. Duffy, D. W. Bednarski, M. Schummer, and D. Haussler, "Support vector machine classification and validation of cancer tissue samples using microarray expression data," Bioinformatics, vol. 16, no. 10, pp. 906–914, 2000.
[47] H. Liu and R. Setiono, "Feature selection via discretization," IEEE Transactions on Knowledge and Data Engineering, vol. 9, no. 4, pp. 642–645, 1997.
[48] H. Vafaie and I. F. Imam, "Feature selection methods: Genetic algorithms vs. greedy-like search," in Proceedings of the International Conference on Fuzzy and Intelligent Control Systems, vol. 51, 1994.
[49] Z. Michalewicz, Genetic algorithms+ data structures= evolution programs. Springer Science & Business Media, 2013.
[50] X. Zhang, Q. Zhang, M. Chen, Y. Sun, X. Qin, and H. Li, "A two-stage feature selection and intelligent fault diagnosis method for rotating machinery using hybrid filter and wrapper method," Neurocomputing, vol. 275, pp. 2426–2439, 2018.
[51] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," IEEE Transactions on Evolutionary Computation, vol. 20, no. 4, pp. 606–626, 2015.
[52] M. Rogati and Y. Yang, "High-performing feature selection for text classification," in Proceedings of the Eleventh International Conference on Information and Knowledge Management, 2002, pp. 659–661.
[53] A. W. Moore and D. Zuev, "Internet traffic classification using Bayesian analysis techniques," in ACM SIGMETRICS Performance Evaluation Review, vol. 33, no. 1, 2005, pp. 50–60.
[54] L. Jun, Z. Shunyi, L. Shidong, and X. Ye, "P2P traffic identification technique," in Proceedings of the 2007 International Conference onComputational Intelligence and Security, 2007, pp. 37–41.
[55] Y.-x. Yang, R. Wang, Y. Liu, and X.-y. Zhou, "Solving P2P traffic identification problems via optimized support vector machines," in Proceedings of the IEEE/ACS International Conference on Computer Systems and Applications (AICCSA'07.
[56] A. F. AlHarthi, "Designing an accurate and efficient classification approach for network traffic monitoring," Ph.D. dissertation, School of Computer Science and Information Technology, RMIT, 2015.
[57] B. M. A. Abdalla, M. Hamdan, M. S. Mohammed, J. S. Bassi, I. Ismail, and M. N. Marsono, "Impact of packet inter-arrival time features for online peer-to-peer (p2p) classification," International Journal of Electrical and Computer Engineering (IJECE), vol. 8, no. 4, pp. 2521–2530, 2018.
[58] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 2, pp. 153–158, 1997.
[59] T. M. Cover and J. A. Thomas, Elements of Information Theory. New York: Wiley, 1991.
[60] M. AlSabah, K. Bauer, and I. Goldberg, "Enhancing Tor's performance using real-time traffic classification," in Proceedings of the 2012 ACM conference on Computer and Communications Security, 2012, pp. 73–84.
[61] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," ACM SIGKDD Explorations Newsletter, vol. 11, no. 1, pp. 10–18, 2009.
[62] A. Montieri, D. Ciuonzo, G. Bovenzi, V. Persico, and A. Pescapé, "A dive into the dark web: Hierarchical traffic classification of anonymity tools," IEEE Transactions on Network Science and Engineering, vol. 7, no. 3, pp. 1043–1054, 2020.

BUSHRA MOHAMMED received a B.Sc. and M.Sc. degree in Computer Engineering and Networks from the University of Gezira, Sudan, and a Ph.D. degree in Electrical Engineering from Universiti Teknologi Malaysia, in 2020. He is a lecturer at the Faculty of Computer and Statistics Studies, University of Kordofan, Sudan. His research interests include computer architecture, network traffic classification and control, artificial intelligence, and optimization techniques.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3037492, IEEE Access

MOSAB HAMDAN received a B.Sc. degree in Computer and Electronic System Engineering from the University of Science and Technology (Sudan) in 2010. He received an M.Sc. in Computer Architecture and Networking from the University of Khartoum (Sudan) in 2014. He is currently pursuing his Ph.D. degree in the School of Electrical Engineering, Faculty of Engineering, Universiti Teknologi Malaysia. His current research interests are software-defined networking (SDN), load balancing, network traffic classification, and future network.

SB JOSEPH received his Ph.D. degree in Electrical Engineering from Universiti Teknologi Malaysia in 2017, M.Eng. degree in Electrical and Electronics Engineering (Electronics) from the University of Maiduguri, Nigeria in 2012, and B.Tech. degree in Computer Science and Mathematics from the Federal University of Technology Minna, Nigeria in 2000. He is currently a Lecturer with the Department of Computer Engineering, Faculty of Engineering, University of Maiduguri, Nigeria. His research interests are data mining, network algorithmic, artificial intelligence and optimization techniques, and computer communication networks.

HA JAMIL is an assistant professor at Elimam Elmahdi university. He received a B.Sc. and M.Sc. from the University of Gezira Sudan and a Ph.D. from Universiti Teknologi Malaysia. His research interests include computer networks, network traffic classification, peer-to-peer computing, and optimization techniques.

SULEMAN KHAN received the Ph.D. degree (Distinction) from the Faculty of Computer Science and Information Technology, University of Malaya, Malaysia, in 2017. He was a faculty member with the School of Information Technology, Monash University Malaysia (June 17-March 19). Currently, he is a faculty member in the Department of Computer and Information Sciences, Northumbria University, Newcastle, UK. He has published more than 50 high-impact research articles in reputed international journals and conferences. His research areas include, but are not limited to, network forensics, software-defined networks, the Internet-of-things, cloud computing, and vehicular communications.

ABDALLAH ELHIGAZI received a B.Sc. degree in Computer Science from the University of Science and Technology (Sudan) in 2004. He received his M.Sc. degree in Computer Science from the University of Gezira (Sudan) in 2007. He received a Ph.D. degree in 2020 in the School of Computing, Faculty of Engineering, Universiti Teknologi Malaysia. His current research interests in wireless network security, machine learning artificial intelligence, information assurance, and security research.

DANDA B. RAWAT is a professor in the Department of Electrical Engineering and Computer Science (EECS), Founding Director of the Howard University Data Science and Cybersecurity Center, Director of Cyber-security and Wireless Networking Innovations (CWiNs) Research Lab, Graduate Program Director of Howard-CS Graduate Programs and Director of Graduate Cybersecurity Certificate Program at Howard University, Washington, DC, USA. Dr. Rawat is engaged in research and teaching in the areas of cybersecurity, machine learning and wireless networking for emerging networked systems including cyber-physical systems, Internet-of-Things, smart cities, software defined systems and vehicular networks. His professional career comprises more than 15 years in academia, government, and industry.

He has secured over $5 million in research funding from US National Science Foundation, US Department of Homeland Security, Department of Energy, National Nuclear Security Administration (NNSA), DoD Research Labs, Industry (Microsoft, Intel, etc.) and private Foundations. Dr. Rawat is the recipient of NSF CAREER Award in 2016, Department of Homeland Security (DHS) Scientific Leadership Award in 2017, the US Air Force Research Laboratory (AFRL) Summer Faculty Visiting Fellowship in 2017, and Outstanding Research Faculty Award (Award for Excellence in Scholarly Activity) at GSU in 2015, the Best Paper Awards and Outstanding PhD Researcher Award in 2009. He has delivered over 15 Keynotes and invited speeches at international conferences and workshops. Dr. Rawat has published over 200 scientific/technical articles and 9 books. He has been serving as an Editor/Guest Editor for over 30 international journals. He has been in Organizing Committees for several IEEE flagship conferences such as IEEE INFOCOM, IEEE CNS, IEEE ICC, IEEE GLOBECOM, IEEE CCNC, IEEE ICNC, IEEE AINA, and so on. He served as a technical program committee (TPC) member for several international conferences including IEEE INFOCOM, IEEE GLOBECOM, IEEE CCNC, IEEE GreenCom, IEEE AINA, IEEE ICC, IEEE WCNC and IEEE VTC conferences. He served as a Vice Chair of the Executive Committee of the IEEE Savannah Section from 2013 to 2017. Dr. Rawat received the Ph.D. degree from Old Dominion University, Norfolk, Virginia. Dr. Rawat is a Senior Member of IEEE and ACM, a member of ASEE, and a Fellow of the Institution of Engineering and Technology (IET).

ISMAHANI BINTI ISMAIL received a Ph.D. degree in Electrical Engineering from University Teknologi Malaysia, Malaysia, in 2013. She is currently a senior lecturer at University Teknologi Malaysia, Malaysia, under the Department of Electronics and Computer Engineering. She works in network algorithmics, digital system, and FPGA system design.

MN MARSONO is an associate professor in Electronic and Computer Engineering, School of Electrical Engineering, Faculty of Engineering, Universiti Teknologi Malaysia. He obtained his Ph.D. in Electrical and Computer Engineering from the University of Victoria BC Canada in 2007, M.Eng. in Electrical Engineering from Universiti Teknologi Malaysia in 2001, and B.Eng. in Computer Engineering from Universiti Teknologi Malaysia in 1999. His research focuses on specialized hardware architecture and network algorithmics for high-throughput packet and flow processing. He works on dynamically reconfigurable platforms for middlebox, fog and edge computing, software-defined networking, and teletraffic engineering. He also works in domain-specific reconfigurable computing research, focusing on multicore/manycore system-on-chip, network-on-chip, design space exploration, mapping, and prototyping of the homogeneous and heterogeneous manycore SoC.