استخراج ویژگی جهت شناسایی ترافیک شبکه با درنظرگرفتن اثرات اتلاف بسته‌ها

محمدرضا گندمی* و حمید حسن‌پور

گروه هوش مصنوعی، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه صنعتی شاهرود، شاهرود، ایران

چکیده

شناسایی ترافیک شبکه یکی از نیازهای اساسی مدیران جهت کنترل شبکه، برای بهبود کیفیت خدمات‌دهی و حفظ امنیت در شبکه است. یکی از چالش‌های اساسی در روش‌های مبتنی بر تحلیل آماری بسته‌ها، شناسایی ترافیک شبکه، مسأله از دست‌دادن (اتلاف) بسته‌ها است که استفاده از ویژگی‌های آماری در تحلیل ترافیک شبکه را با مشکل جدی روبه‌رو می‌سازد. این مسأله، ویژگی‌های آماری بسته‌ها نظیر فاصله زمانی بین ارسال بسته‌های متوالی برنامه‌های کاربردی را تحت تأثیر قرار می‌دهد، و در مواردی دقت شناسایی ترافیک را به میزان قابل توجهی کاهش می‌دهد. هدف اصلی این مقاله بررسی تأثیرات اتلاف بسته‌ها بر روی ویژگی‌های آماری، و در نتیجه دقت شناسایی برنامه‌های کاربردی، و همچنین استخراج ویژگی‌های مناسب جهت چیره‌شدن بر این تأثیرات است. بدین منظور، رفتار چهار ویژگی آماری، مورد بررسی قرار گرفته و با استخراج ویژگی از توزیع آنها ترافیک شبکه شناسایی می‌شود. به همین منظور پایگاه داده‌ای از ترافیک هفت برنامه کاربردی با نرخ‌های مختلفی از اتلاف بسته، تهیه شده و میزان صحت تشخیص برنامه‌های کاربردی به‌وسیله شبکه عصبی، مورد تحلیل قرار گرفته است. نتایج نشان می‌دهد که ویژگی‌های استخراج‌شده در مقابل رخداد اتلاف بسته‌ها مقاوم بوده و دقت شناسایی ترافیک شبکه را در حالت‌های مختلف رخداد اتلاف بسته به حالت ایده‌آل (عدم رخداد اتلاف بسته در شبکه) نزدیک می‌کند.

واژگان کلیدی: ترافیک شبکه، شناسایی ترافیک شبکه، یادگیری ماشین، از دست دادن بسته

Feature Extraction to Identify Network Traffic with Considering Packet Loss Effects

Mohammadreza Gandomi* & Hamid Hassanpour

Faculty of Computer Engineering and IT, Shahrood University of Technology, Shahrood, Iran.

Abstract

There are huge petitions of network traffic coming from various applications on Internet. In dealing with this volume of network traffic, network management plays a crucial rule. Traffic classification is a basic technique which is used by Internet service providers (ISP) to manage network resources and to guarantee Internet security. In addition, growing bandwidth usage, at one hand, and limited physical capacity of communication lines, at the other hand, lead providers to improve utilization quality of network resources. In fact, classification or identification of network is a critical task in network processing for traffic management, anomaly detection, and also to improve network quality-of-service (QoS). Port and payload based methods are two classical techniques which are applicable under traditional network conditions. However, many Internet applications use dynamic port numbers for communications, which lead to difficulties in identifying traffic using port numbers. Also many applications encrypt the data before transmitting to avoid detection. Therefore, payload-based techniques are inefficient for these traffics. In recent years, statistical feature-based traffic flow identification methods (STFIM) have attracted the interest of many researchers. The most important part of a STFIM is the selection of efficient statistical features.

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

● تاریخ ارسال:۱۳ ۱۰ ۹۶ ● تاریخ پذیرش: ۰۳/۰۲ ۹۷ ● تاریخ انتشار: ۲۸ ۱۲ ۹۸

[DOI: 10.29252/jsdp.16.4.3]

Preliminary analysis shows that the problem of packet loss in data transmission is one of the major challenges in employing STFIM for network traffic identification. This affects the statistical characteristics of packets, such as the time interval between sending successive application packets, and in some cases significantly reduces the accuracy of traffic identification. The main goal of this paper is to examine the effects of packet loss on statistical features, and therefore the accuracy of identifying applications, as well as extracting appropriate features to overcome these effects. For this purpose, the behavior of four statistical features, including the packet size, the time interval between sending and receiving packets, the duration of the flows and the rate of sending packets, are investigated; then applications traffics are identified via considering characteristics of their distribution.

We collected a database of network traffic flow from seven applications with different rates of packet loss. We used the extracted features in a multilayer neural network, as a classifier, to differentiate between different traffic applications. Experimental results show that the extracted features are robust against the packets loss, and the accuracy of the network traffic identification is close to the ideal state (traffic flow with no packet lost).

Keywords: Network Traffic, Network traffic Identification, Machine Learning, Packet Loss

۱- مقدمه

در سال‌های اخیر همگام با توسعه اینترنت، برنامه‌های کاربردی در حال اجرا بر روی این بستر از قبیل شبکه‌های اجتماعی، بازی‌های برخط و سرویس‌های بر پایه ابر[1] روز به‌روز عمومیت بیشتری یافته‌اند؛ لذا تقاضای رو به رشد کاربران برای استفاده از پهنای باند بیشتر، و از طرف دیگر محدودیت ظرفیت فیزیکی خطوط ارتباطی شبکه، سرویس‌دهندگان اینترنت را بر آن می‌دارد تا با اولویت‌دهی تخصیص منابع کیفیت بهره برداری کاربران از منابع شبکه را بهبود بخشند. همچنین با توجه به افزایش بدافزارها و تلاش آن‌ها برای پنهان‌سازی ترافیک خود به‌منظور گریز از سامانه‌های تشخیص نفوذ و دیوارهای آتش، شناسایی و طبقه‌بندی ترافیک برنامه‌های کاربردی امری ضروری است. ازاین‌رو، طبقه‌بندی ترافیک شبکه برای انجام بسیاری از وظایف امنیتی، مدیریتی و کنترلی در شبکه لازم و ضروری است.

روش‌های موجود جهت شناسایی و طبقه‌بندی ترافیک برنامه‌های کاربردی را می‌توان در چهار دسته مبتنی بر (۱) شماره درگاه، (۲) بررسی محتوای بسته‌های ترافیک، (۳) رفتار کاربر، و (۴) ویژگی‌های آماری بسته‌های ترافیک در حال عبور، تقسیم‌بندی کرد. در این میان روش‌های مبتنی بر تحلیل آماری بسته‌ها، اهمیت بالایی دارند. در این روش‌ها، عملیات شناسایی ترافیک برنامه‌های کاربردی بدون نیاز به بررسی محتوای بسته‌های ترافیک و با هدف حفظ محرمانگی اطلاعات مبادله‌شده انجام می‌شود [17,18].

یکی از چالش‌های اصلی در روش‌های مبتنی بر ویژگی‌های آماری که کمتر در مقالات به آن پرداخته شده است، تأثیرپذیری آنها از رخداد ازدحام بسته‌ها و اتلاف بسته[2] در هنگام انتقال در بستر شبکه است [1]. به‌طورعمومی اتلاف بسته هنگام ازدحام در شبکه و عدم توانایی مسیریاب در رسیدگی به تمامی بسته‌های دریافتی، و دورریختن آنها از صف، رخ می‌دهد. پژوهش‌های بسیاری در زمینه جلوگیری از رخداد اتلاف بسته و بازیابی بسته‌ها به جهت افزایش کیفیت سرویس‌دهی انجام شده است [2,3]. در هنگام وقوع اتلاف بسته، دنباله‌ای از بسته‌ها از دست می‌روند، و درنتیجه ویژگی‌های آماری مورد استفاده جهت شناسایی برنامه کاربردی نیز تغییر می‌کنند. از جمله این تغییرات می‌توان به تغییر زمان ارسال و دریافت بسته‌ها، تغییر در ترتیب بسته‌ها، تغییر در نرخ ارسال و دریافت بسته‌ها و مدت زمان جریان‌ها اشاره کرد. در سال‌های اخیر روش‌های متعددی جهت شناسایی ترافیک شبکه ارائه شده است، ولی بررسی‌های ما نشان می‌دهد که تاکنون راه‌حلی برای موضوع رخداد اتلاف بسته در شناسایی ترافیک شبکه ارائه نشده است و روش‌های ارائه‌شده جهت شناسایی ترافیک شبکه بر روی پایگاه داده استاندارد (بدون اتلاف بسته) مورد ارزیابی قرار گرفته‌اند [4, 15, 16].

با توجه به شایع‌بودن رخداد اتلاف بسته در شبکه‌های رایانه‌ای، هدف این مقاله بررسی رخداد اتلاف بسته، و تأثیر آن بر ویژگی‌های آماری و دقت شناسایی برنامه‌های کاربردی است و در ادامه کار روشی جهت استخراج ویژگی‌های آماری مقاوم در برابر رخداد اتلاف بسته ارائه می‌شود. به همین منظور پایگاه داده‌ای از هفت برنامه کاربردی شامل کروم، فایرفاکس، اینترنت اکسپلورر، اسکایپ، تلگرام، دانلود منیجر و توییتر با حالت‌های مختلف اتلاف بسته فراهم شده است تا تأثیر اتلاف بسته بر روی ویژگی‌ها و دقت شناسایی بررسی شود. دلیل انتخاب این برنامه‌های کاربردی را می‌توان عمومیت آنها در استفاده‌های روزانه کاربران شبکه دانست و همچنین شناسایی

[1] Cloud
[2] Packet Loss

[DOI: 10.29252/jsdp.16.4.3]

این برنامه‌های کاربردی یکی از نیازهای ضروری مدیران شبکه‌ها است. در ادامه کار با تحلیل فراوانی توزیع داده‌ها در چهار ویژگی اندازه بسته‌ها، فاصله زمانی بین بسته‌های ارسالی و دریافتی، مدت‌زمان جریان بسته‌ها و نرخ ارسال و دریافت بسته‌ها، ویژگی‌هایی مقاوم به رخداد اتلاف بسته، جهت شناسایی برنامه‌های کاربردی استخراج می‌شود؛ سپس با اعمال این ویژگی‌ها به یک شبکه عصبی، عمل شناسایی ترافیک شبکه انجام می‌شود. نتایج حاصل از بررسی روش ارائه‌شده نشان می‌دهد که با استفاده از این ویژگی‌های استخراج‌شده، دقت شناسایی برنامه‌های کاربردی هنگام رخداد اتلاف بسته در شبکه، نزدیک به دقت شناسایی در حالت عدم رخداد اتلاف بسته است.

در ادامه مقاله، در بخش دوم کارهای انجام‌گرفته در زمینه استخراج ویژگی از ترافیک برنامه‌های کاربردی و روش‌های موجود برای شناسایی آنها مرور می‌شود. در بخش سوم، درخصوص پایگاه داده جمع‌آوری‌شده با شرایط مختلف اتلاف بسته و تأثیرات آن بر ویژگی‌های آماری و دقت شناسایی برنامه‌های کاربردی بحث می‌شود. در بخش چهارم روش پیشنهادی ارائه می‌شود و با توجه به این روش، اتلاف بسته با نرخ‌های مختلف مورد بررسی و ارزیابی قرار می‌گیرد. در بخش پنجم، نتایج حاصل از به‌کارگیری روش پیشنهادی جهت شناسایی ترافیک برنامه‌های کاربردی بررسی و در بخش ششم به جمع‌بندی مسائل پرداخته می‌شود.

۲- مروری بر کارهای گذشته

به‌طور کلی عمده کارهای انجام‌شده در حوزه شناسایی ترافیک شبکه بر پایه ویژگی‌های آماری را می‌توان به دو گروه تقسیم‌بندی کرد. در گروه نخست، یک یا چند ویژگی جدید از بسته‌ها و جریان‌ها برای تمایز بین ترافیک برنامه‌های کاربردی مختلف استخراج می‌شود. در گروه دوم، با استفاده از ویژگی‌های مرسوم مورد استفاده در مقالات، و به‌کارگیری یک روش یادگیری نوین، دسته‌بندی برنامه‌های کاربردی انجام می‌شود.

Kim و *همکارانش* روشی برای طبقه‌بندی ترافیک ارائه کردند که بر پایه مدل مارکف عمل کرده و از معیار هم‌گرایی Kullback-Leibler (یک معیار شباهت در بین توزیع‌های احتمالاتی) جهت بررسی ترافیک برنامه‌های کاربردی استفاده شده است. این روش با دقت ۹۰٪ قادر به شناسایی ترافیک برنامه‌های کاربردی در شرایط تداخل بسته‌ها است [5]. در این روش از یادگیری بانظارت استفاده شده و حالت‌های مدل مارکف به‌وسیله جهت انتقال و اندازه چهار بسته نخست ساخته می‌شود. Muehlstein و *همکارانش* نشان دادند که با بررسی ترافیک شبکه نوع سیستم‌عامل، نوع مرورگر و نوع برنامه کاربردی داده‌های HTTP و داده‌های رمزشده HTTPS را می‌توان تعیین کرد [6]. نویسندگان ادعا کردند که در ترافیک رمزشده، برای نخستین‌بار موفق به شناسایی سیستم‌عامل، مرورگر و برنامه کاربردی شدند. در آن پژوهش، پایگاه داده‌ای حاوی بیست‌هزار نشست، مورد ارزیابی قرار گرفت؛ که شامل ترافیک سیستم‌عامل‌های ویندوز، اوبونتو، آی‌او‌اس (IOS) و مرورگر های کروم، اینترنت اکسپلورر، فایرفاکس و سافاری، و برنامه‌های کاربردی یوتیوب، فیسبوک و توئیتر بوده است. در این روش، هر نشست که شامل پنج‌تایی (شماره درگاه مبداء و مقصد، نشانی مبداء و مقصد، پروتکل مربوطه) است، به یک سه‌تایی (سیستم‌عامل، مرورگر و برنامه کاربردی) نگاشت می‌شود. در این مقاله از الگوریتم یادگیری SVM و هسته RBF استفاده شده است.

Loo و *همکارانش* الگوریتمی بر پایه خوشه‌بندی k-means افزایشی جهت یادگیری نمونه‌های دارای برچسب و بدون برچسب ارائه داده‌اند و جهت سنجش میزان شباهت نمونه‌ها از دو معیار فاصله اقلیدسی و منهتن استفاده کرده است. آزمایش بر روی ۶۷۱ هزار جریان انجام شده و میزان دقت خوشه‌بندی برابر ۹۴ درصد گزارش شده است. همچنین سرعت اجرای الگوریتم با استفاده از معیار فاصله منهتن سه برابر بهتر از فاصله اقلیدسی ارزیابی شده است [7]. Qin و *همکارانش* مدلی به نام مدل "جریان دوطرفه" ارائه داده‌اند که می‌تواند خصوصیات رفتاری متقابل بین نودهای مختلف را ضبط کند. از توزیع احتمالی اندازه محتوای هر بسته[1] (PSD) موجود در این مدل به‌عنوان ویژگی استفاده شده است [8]. بسته‌های (رفت و برگشت) دارای مبدأ و مقصد یکسان به‌عنوان یک جریان دوطرفه در نظر گرفته می‌شوند. در مرحله بعد نشان داده شده که توزیع اندازه بسته‌های مربوط به برنامه‌های کاربردی مختلف بایکدیگر متفاوت است؛ سپس از Renyi Cross Entropy برای محاسبه شباهت بین PSD یک جریان دو طرفه با PSD برنامه‌های کاربردی شناخته‌شده استفاده می‌شود.

علی‌اکبریان و همکاران با استفاده از الگوریتم‌های یادگیری ماشین نظارت‌شده سعی در نشان‌دادن بهترین تعداد ویژگی برای شناسایی ترافیک داشته‌اند [9]؛ همچنین از الگوریتم‌های درخت تصمیم، شبکه عصبی مصنوعی، یادگیری

[1] Packet Size Distribution

[DOI: 10.29252/jsdp.16.4.3]

بیز و Bagging و Boosting جهت طبقه‌بندی ترافیک استفاده کرده‌اند. بر اساس نتایج این مقاله، اگر تعداد M رده برنامه کاربردی داشته باشیم، تعداد ویژگی‌ها در بهترین حالت می‌تواند به M-1 کاهش یابد. Yamansavascilar و *همکارانش* سعی داشتند که برنامه‌های کاربردی همچون توییتر، فیسبوک و اسکایپ بر روی دو پایگاه داده UNB و پایگاه دادهٔ جمع‌آوری‌شده حاوی تعدادی برنامه کاربردی شناسایی کنند [4]. برای ارزیابی نتایج از چهار الگوریتم دسته‌بندی J48، Random forest، K-NN و Bayesnet استفاده کردند. نتایج حاصل از انجام آزمایش توسط الگوریتم K-NN با مجموعه ۱۱۱ ویژگی برابر ۹۴/۹۳٪ و با استفاده از Random Forest دقت شناسایی برابر ۸۷ ۹۰٪ شده است.

جهت شناسایی مؤثر ترافیک شبکه به‌صورت برخط روشی با استفاده از درخت شبکه عصبی انعطاف‌پذیر (FNT[1]) و استفاده از سه مجموعه داده جهت ارزیابی این روش ارائه شده است [10]. در این روش از ویژگی‌های اندازه بسته‌ها و میانگین آنها، بیشترین مقدار، کمترین مقدار و انحراف معیار اندازه بسته‌ها در جریان‌ها استفاده شده است، و توانست با بررسی شش بسته ابتدایی برنامه کاربردی را شناسایی کند. Ertam و *همکارانش* برای شناسایی ترافیک شبکه از روش ELM[2] استفاده کرده‌اند. در این روش ابتدا از شبکه عصبی با یک لایهٔ مخفی استفاده شده که در آن وزن نرون‌های ورودی و مخفی به‌صورت تصادفی تعیین می‌شود. میزان دقت دسته‌بندی توسط تغییر تعداد لایه‌های مخفی و هم‌چنین تغییر پارامتر تابع هدف (تابع موجک) به ۹۵ ٪ رسیده است [11].

همان‌طور که مشاهده می‌شود، کارهای متعددی در سال‌های اخیر در حوزه شناسایی ترفیک شبکه انجام شده است، ولی به‌طورعمومی این روش‌ها به ارزیابی نتایج بر روی پایگاه داده استاندارد پرداختند و برخی تأکید کرده‌اند که مجموعه دادهٔ جمع آوری‌شده و مورد ارزیابی قرارگرفته حاوی اتلاف بسته نیست [12]؛ ازاین‌رو رخداد اتلاف بسته در شبکه یکی از چالش‌های اصلی در شناسایی ترافیک شبکه بوده و با توجه به پژوهش‌های انجام‌شده، روشی مؤثر جهت حل این مسأله ارائه نشده است.

۳- اتلاف بسته و تأثیر آن بر ویژگی‌های آماری

در یکی از پژوهش‌های اخیر، با استفاده از تحلیل رفتار برنامه‌های کاربردی، پنج ویژگی آماری (ترتیب بسته‌ها، اندازه بسته‌ها، مدت زمان مابین ارسال و دریافت بسته‌ها، مدت زمان جریان‌ها و نرخ ارسال بسته‌ها) از بسته‌های ترافیک شبکه استخراج شد. به‌منظور ارزیابی این روش، شناسایی ترافیک شش برنامه کاربردی بر روی دو پایگاه داده UNIBS و پایگاه داده جمع‌آوری‌شده با خصوصیت عدم رخداد اتلاف بسته انجام گرفت. این روش با استفاده از الگوریتم Random Forest و این پنج ویژگی به دقت ۹۷/۵ درصد جهت شناسایی این شش برنامه کاربردی رسیده است [13].

همان‌طور که درقبل اشاره شد، نکته اصلی و چالش‌برانگیز در این کار و همچنین مقالات اخیر در حوزه شناسایی ترافیک شبکه، رخداد اتلاف بسته است. ازدست‌دادن بسته‌ها هنگامی رخ می‌دهد که یک یا چند بسته از داده‌های ارسال‌شده در بستر شبکه رایانه‌ای به مقصد نرسد [14]. به‌طورمعمول اتلاف بسته در اثر ازدحام در شبکه رخ می‌دهد و با پروتکل TCP قابل تشخیص است و عملیات ارسال مجدد بسته‌ها توسط این پروتکل جهت برقراری قابلیت اطمینان و افزایش کارایی در پیام‌رسانی انجام می‌شود. در شبکه‌ای که با اتلاف بسته همراه است پروتکل TCP سازوکار ارسال و دریافت بسته‌ها در شبکه را با استفاده از پنجره ارسال تغییر می‌دهد و بسته‌های ازدست‌رفته را دوباره با ترتیب خاصی ارسال می‌کند؛ ازاین‌رو در این بخش تأثیر اتلاف بسته بر ویژگی‌های آماری و دقت شناسایی ترافیک برنامه‌های کاربردی بررسی می‌شود. قدم نخست در رسیدن به این هدف تهیه پایگاه داده‌ای از ترافیک شبکه است که در آن، اتلاف بسته رخ داده باشد.

۳-۱- پایگاه داده

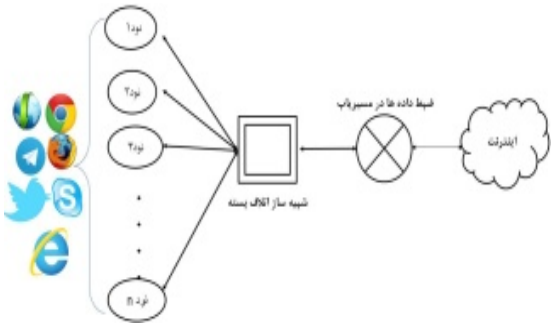
جهت بررسی تأثیرات اتلاف بسته (به جهت عدم وجود پایگاه داده) لازم بود تا مجموعه‌داده‌ای از ترافیک برنامه‌های کاربردی مختلف ضبط شود که در آن اتلاف بسته با نرخ مختلفی رخ داده باشد؛ ازاین‌رو با استفاده از معماری شکل (۱) به شبیه‌سازی رخداد اتلاف بسته در شبکه پرداخته می‌شود که در آن برنامه‌های کاربردی در نودهای مختلف شبکه در حال اجرا بوده و بسته‌های مربوطه در بستر شبکه ارسال می‌شوند.

روش کار به این صورت است که سامانه شبیه‌سازی رخداد اتلاف بستهٔ ما بین ارسال و دریافت بسته‌ها در شبکه قرار می‌گیرد و بسته‌های ارسالی و دریافتی درون صف مربوط به این سامانه وارد می‌شوند. در این شبیه‌سازی نرخ ورودی به صف برابر نرخ خروجی صف نبوده و در هنگام رسیدگی به بسته‌های ورودی به صف، پدیده ازدحام یا سرریزی صف رخ

[1] Flexible Neural Trees
[2] Extreme Learning Machine

[DOI: 10.29252/jsdp.16.4.3]

می‌دهد، و توسط تابعی بسته‌ها به‌صورت به‌طورکامل تصادفی و با نرخ تنظیم‌شده دور ریخته می‌شوند. به این ترتیب پروتکل TCP در مبدا ارسال بسته‌ها درصورتی‌که بسته‌های برگشت را دریافت نکرده باشد، سرعت ارسال را کاهش داده و بسته‌های ارسالی را دوباره ارسال می‌کند؛ ازاین‌رو در شبکه اتلاف بسته رخ می‌دهد و بسته‌هایی که ضبط می‌شوند، تحت تأثیر عملکرد پروتکل TCP جهت ارسال مجدد بسته‌ها قرار می‌گیرند. داده‌های ضبط‌شده در اثر رخداد اتلاف بسته در شبکه طی آزمایش‌های مختلفی برای نرخ‌های اتلاف بسته صفر، ۱۰، ۲۰، ۲۵، ۳۰، ۴۰، ۵۰ و ۷۰ درصد انجام شده است. جدول (۱) تعداد بسته‌های ضبط‌شده از هر برنامه کاربردی با درصدهای مختلف اتلاف بسته را نشان می‌دهد. این برنامه‌های کاربردی جزئی از اصلی‌ترین برنامه‌های مورد استفاده به‌صورت مستمر هستند؛ همچنین سعی شده است تا از گروه‌های مختلف (نظیر به نظیر، مشتری- خدمت‌گذار و ابزار بارگذاری) ترافیک تولید و ضبط شود.



(شکل-۱): معماری سامانه جهت شبیه‌سازی رخداد اتلاف بسته و ضبط داده‌ها

(Figure-1): System Structure to Emulate Packet Loss and Data Capturing

۲-۳- شناسایی برنامه کاربردی

پس از تهیه پایگاه داده، عملیات شناسایی برنامه‌های کاربردی با استفاده از الگوریتم Random Forest انجام می‌شود (جزئیات انجام این‌کار در [13] آمده است). برای انجام این‌کار، داده‌های فاقد اتلاف بسته جهت اموزش، و داده‌های با ۱۰، ۲۰، ۵۰ و ۷۰ درصد اتلاف بسته جهت آزمایش، مورد استفاده قرار گرفت. نتایج شناسایی هفت برنامه کاربردی فایرفاکس، کروم، اینترنت اکسپلورر، اسکایپ، تلگرام، دانود منیجر و توئیتر در شکل (۲) نشان داده شده است.

در شکل (۲) نمودار مقادیر دقت[1] و بازخوانی[2] به‌دست‌آمده از این آزمایش‌ها نشان داده شده است. همان‌طورکه مشاهده می‌شود دقت شناسایی با رشد نرخ اتلاف بسته کاهش پیدا می‌کند؛ دلیل این امر را می‌توان این‌طور بیان کرد که با ازدست‌دادن بسته‌ها و ارسال مجدد آنها توسط پروتکل TCP، رفتار داده‌ها نیز تغییر می‌کند. برای نشان‌دادن رفتار داده‌ها و میزان تغییر رفتارهای نرخ‌های مختلف اتلاف بسته از تابع توزیع تجمعی (CDF[3]) می‌توان استفاده کرد. نمودار توزیع تجمعی چهار ویژگی اندازه بسته‌ها، مدت زمان بین ارسال و دریافت بسته‌ها، مدت زمان جریان‌ها و نرخ ارسال بسته‌ها برای هفت برنامه کاربردی با نرخ‌های مختلف اتلاف بسته در شکل (۳) نشان داده شده است. در کار قبلی، برنامه‌های کاربردی به دو دسته Client-Server و P2P تقسیم‌بندی شدند و نشان داده شد که رفتار ویژگی‌های آماری این دو دسته در داخل دسته‌ها به یکدیگر شبیه بوده و این رفتار بین این دو دسته به‌طورکامل متفاوت است [13]. طبق بررسی‌های به‌عمل‌آمده، در برنامه‌های Client-Server هنگامی که میزان رخداد اتلاف بسته از پنجاه درصد افزایش یابد، کارایی برنامه از بین رفته و صفحه مربوط به برنامه بارگذاری نمی‌شود.

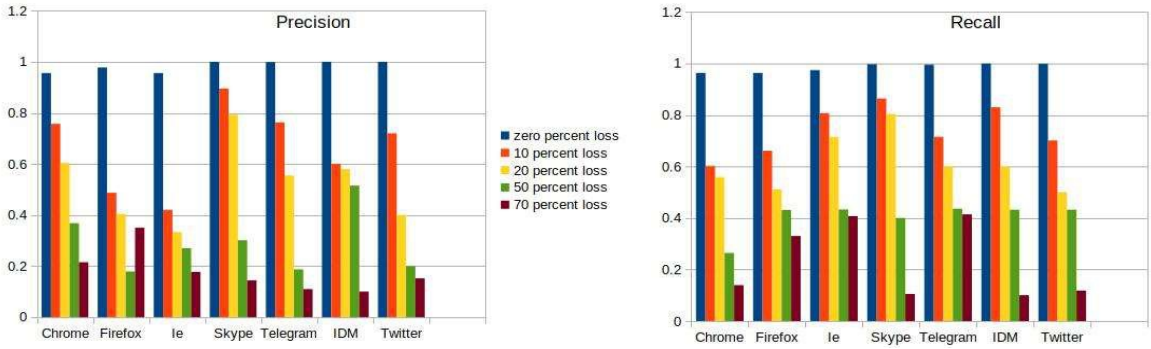
(جدول-۱): پایگاه داده جمع‌آوری‌شده با رخدادهای مختلف اتلاف بسته

(Table-1): Collected Database with Different Packet Loss Rate

برنامه کاربردی	صفر درصد	٪۱۰	٪۲۰	٪۲۵	٪۳۰	٪۴۰	٪۵۰	٪۷۰
Chrome	۲۴۵۳۵۹	۲۴۳۰۳۲	۲۳۳۱۲۴	۲۳۲۲۲۴	۲۳۵۹۵۰	۲۳۲۲۸۶	۲۳۵۹۵۰	۲۳۵۵۸۸
Firefox	۲۴۴۰۳۲	۲۱۱۱۲۴	۲۴۱۰۳۲	۲۱۷۶۷۴	۲۷۷۰۳۲	۲۴۰۳۴۱	۲۴۵۰۳۲	۲۳۲۵۸۸
Internet Explorer	۲۹۷۶۷۴	۲۳۱۳۵۹	۲۶۴۰۳۲	۲۳۳۶۷۴	۲۴۳۹۴۱	۲۷۰۵۸۸	۲۳۴۵۸۸	۲۴۹۴۴۲
Skype	۲۱۹۰۳۲	۲۲۹۰۳۲	۲۱۳۳۵۹	۲۴۹۴۳۲	۲۶۰۵۶۶	۲۲۵۶۹۱	۲۲۷۷۹۱	۲۵۵۲۸۶
Telegram	۲۸۷۷۹۱	۲۳۷۷۲۴	۲۳۵۳۵۹	۲۶۲۴۶۶	۲۲۱۷۹۱	۲۰۰۵۸۸	۲۹۲۲۸۶	۲۸۵۹۵۰
IDM	۲۴۱۱۴۱	۲۷۷۴۶۶	۲۳۵۲۲۴	۲۱۸۸۵۹	۲۱۲۷۹۱	۲۹۹۰۳۲	۱۹۲۲۸۶	۲۴۵۵۳۲
Twitter	۲۹۸۹۴۱	۲۵۸۹۵۰	۲۲۲۴۶۶	۲۶۹۱۶۶	۲۲۵۳۵۹	۲۱۳۵۵۹	۳۱۲۲۸۶	۲۱۱۰۳۲

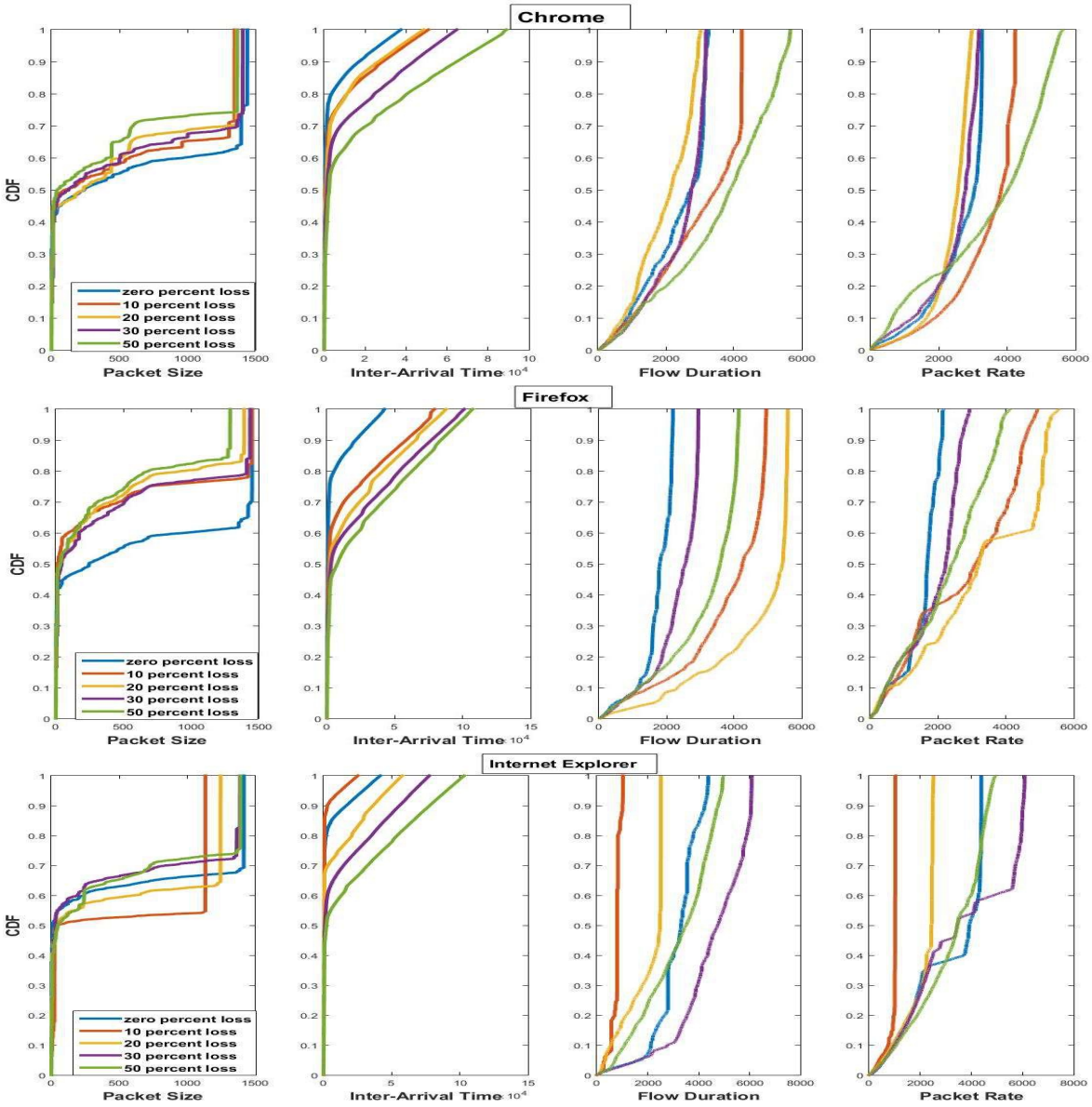
[1] Precision [2] Recall [3] Cumulative Distribution Function

[DOI: 10.29252/jsdp.16.4.3]

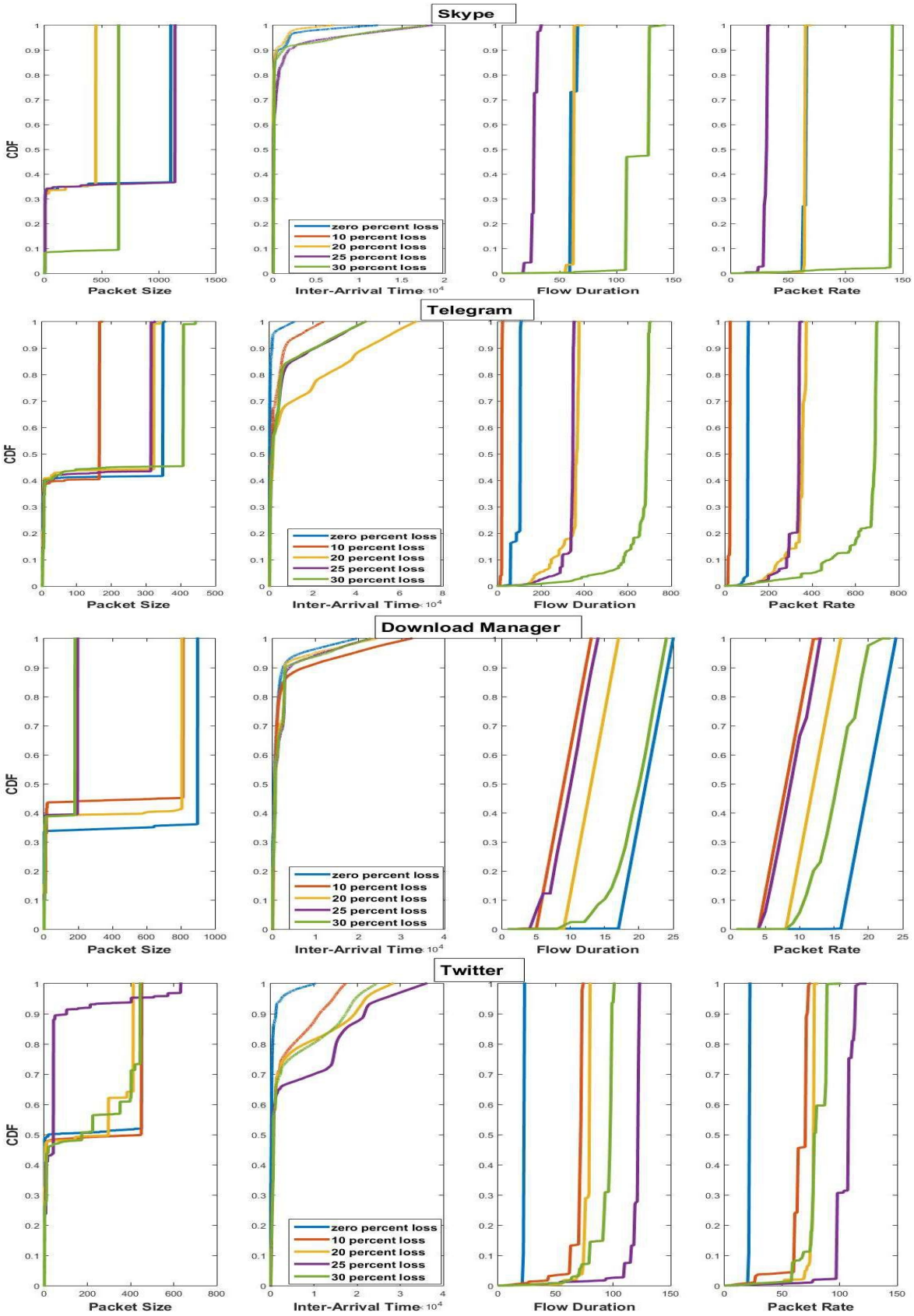


(شکل-۲): نتایج به‌دست‌آمده از شناسایی برنامه‌های کاربردی مختلف از پایگاه داده با درصدهای مختلف اتلاف بسته

(Figure-2): Results of Application Identification with Different Packet Loss Rate



[DOI: 10.29252/jsdp.16.4.3]



(شکل-۳): نمودار توزیع تجمعی چهار ویژگی آماری مربوط به هفت برنامه‌کاربردی با نرخ‌های مختلف اتلاف بسته

(Figure-3): CDF of Four Statistical Feature for Seven Application with Different of Packet Loss Rate

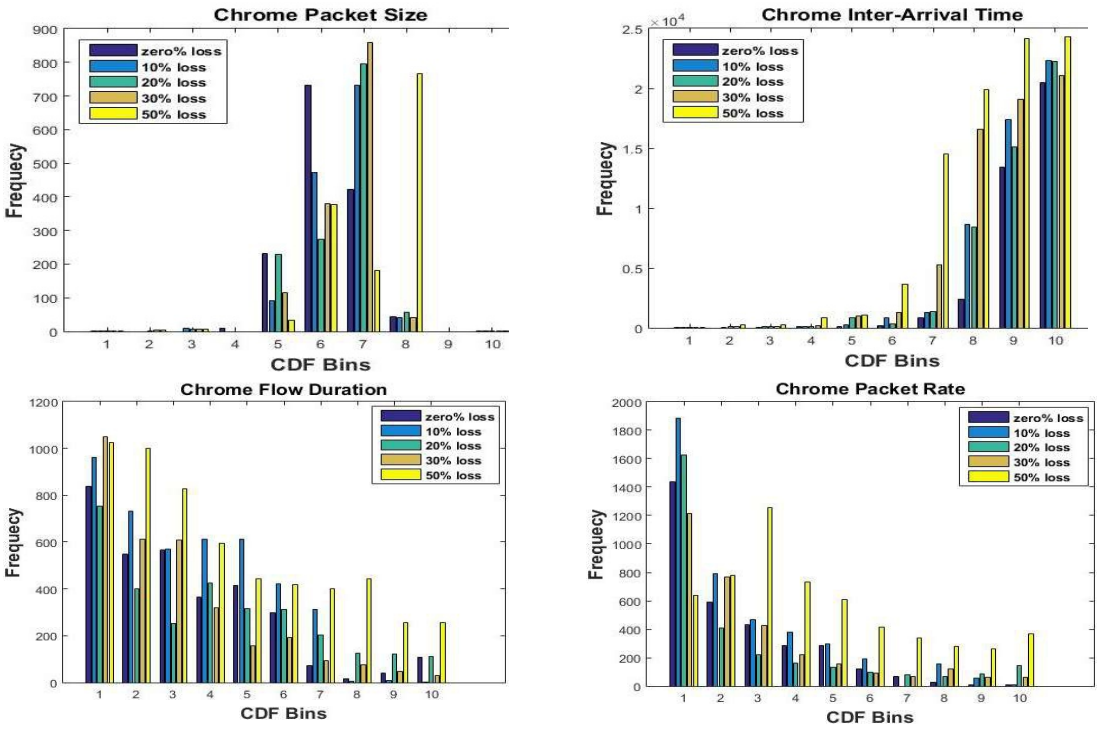
[DOI: 10.29252/jsdp.16.4.3]

این میزان اتلاف بسته در برنامه‌های P2P به میزان سی درصد کاهش می‌یابد. به بیانی دیگر در برنامه‌های کاربردی P2P هنگامی که میزان اتلاف بسته به بیش از سی درصد افزایش یابد پیام‌های به‌صورت صوت یا تصویر بارگذاری نمی‌شوند. به همین دلیل رفتار برنامه‌های Client-Server در بازه صفر تا پنجاه درصد اتلاف بسته و برای برنامه‌های P2P بازه رخداد اتلاف بسته بین صفر تا سی درصد مورد بررسی قرار می‌گیرد.

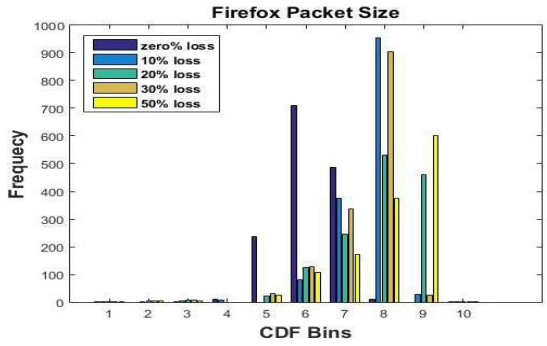
در شکل (۳) توزیع تجمعی داده‌های چهار ویژگی آماری مربوط به بسته‌های هفت برنامه کاربردی نشان داده شده است. همان‌طور که در این شکل قابل مشاهده است، توزیع داده‌ها در تمامی این ویژگی‌ها همزمان با رشد نرخ اتلاف بسته تغییر کرده و این امر موجب کاهش دقت شناسایی برنامه‌های کاربردی می‌شود. به همین منظور در بخش بعد روشی برای استخراج ویژگی‌های مقاوم و موثر در مقابل اتلاف بسته ارائه می‌شود تا بتوان دقت شناسایی برنامه‌های کاربردی را در محیط‌های واقعی که اغلب با اتلاف بسته همراه هستند، به دقت شناسایی در محیط‌های ایده‌آل نزدیک کرد.

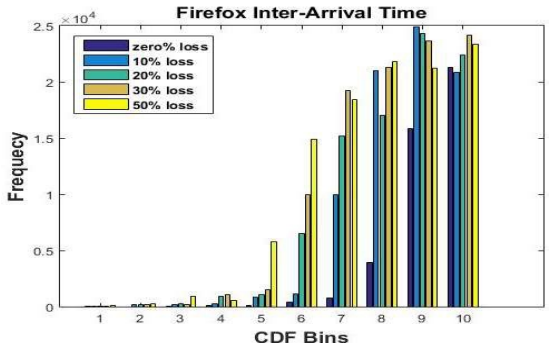
۴- روش پیشنهادی

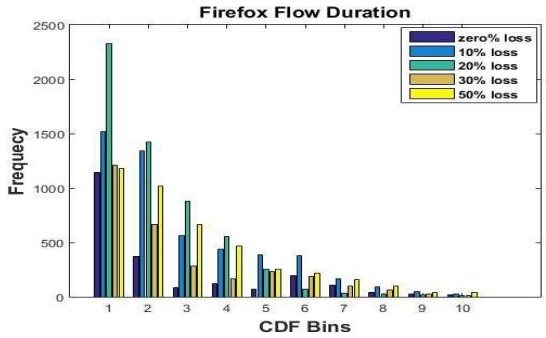
در بخش قبل و در شکل (۳) توزیع تجمعی داده‌های چهار ویژگی آماری در حالت‌های مختلف اتلاف بسته برای هفت برنامه کاربردی نشان داده شد. در آن شکل نمودارهای توزیع تجمعی، با روند افزایشی خود از صفر به یک، به‌خوبی رفتار داده‌ها را نشان می‌دهند. برای تحلیل رفتار داده‌ها، فراوانی تابع توزیع تجمعی در بازه‌های کوتاه و یکسان را مورد بررسی قرار می‌دهیم. در این مقاله با بهره‌گیری از این ویژگی -فراوانی تابع توزیع تجمعی- رفتار داده‌های ترافیکی برنامه‌های کاربردی مختلف مورد بررسی قرار می‌گیرند. در شکل (۴) فراوانی تابع توزیع تجمعی داده‌های مربوط به چهار ویژگی اندازه بسته‌ها، مدت زمان بین ارسال و دریافت بسته‌ها، مدت زمان جریان‌ها و نرخ ارسال بسته‌ها برای هفت برنامه کاربردی با نرخ مختلف اتلاف بسته در بازه‌های کوتاه ۰/۱ نشان داده شده است. همان‌طور که مشاهده می‌شود، نمودار فراوانی مقادیر تابع توزیع تجمعی برای هر یک از برنامه‌های کاربردی با نرخ مختلف اتلاف بسته به یکدیگر شبیه هستند. و در برنامه‌های کاربردی مختلف به‌خوبی قابل تمییز هستند. به‌عنوان نمونه نمودار توزیع داده‌های برنامه کاربردی Skype برای سه ویژگی اندازه بسته، مدت‌زمان بین ارسال و دریافت بسته‌ها و مدت‌زمان جریان‌ها را در نظر بگیرید. نمودار فراوانی هر یک از این ویژگی‌ها با نرخ مختلف اتلاف بسته، شباهت‌های بسیار زیادی با یکدیگر دارند؛ ولی اگر همین نمودارهای فراوانی را با نمودارهای فراوانی تابع توزیع برنامه کاربردی Chrome مقایسه کنیم، اختلاف فاحشی دارند. گفتنی است که بردار ویژگی استخراج‌شده دارای طول ثابت ده است و با به‌کارگیری آن در شبکه عصبی به‌راحتی می‌توان برنامه‌های کاربردی مختلف را از هم تمییز داد.

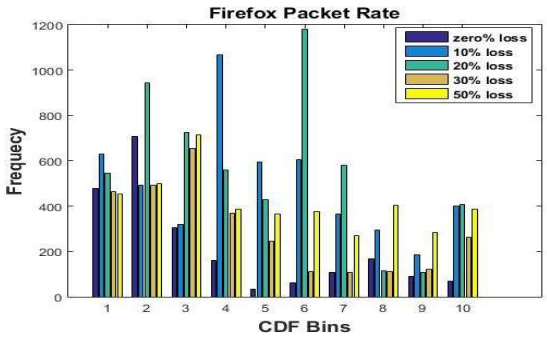


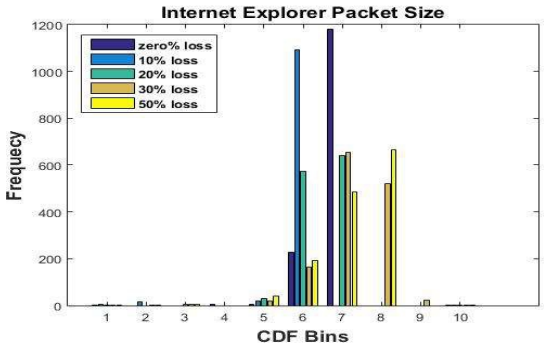
[DOI: 10.29252/jsdp.16.4.3]

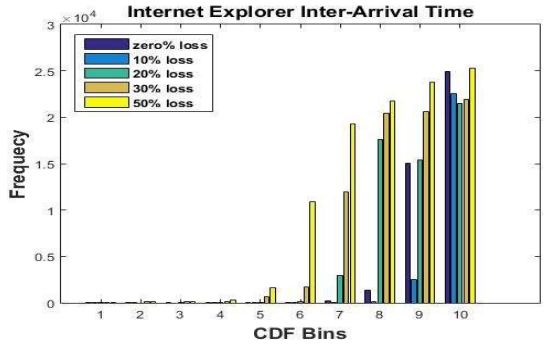


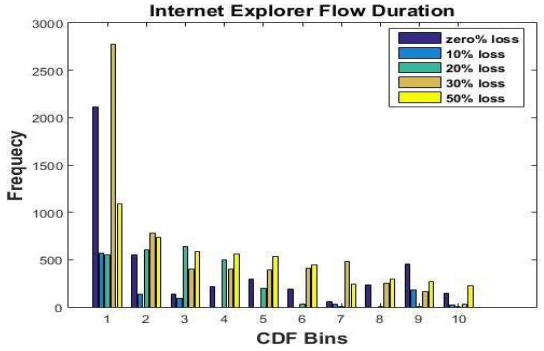


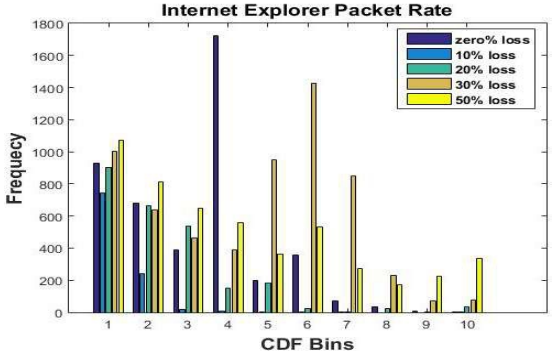


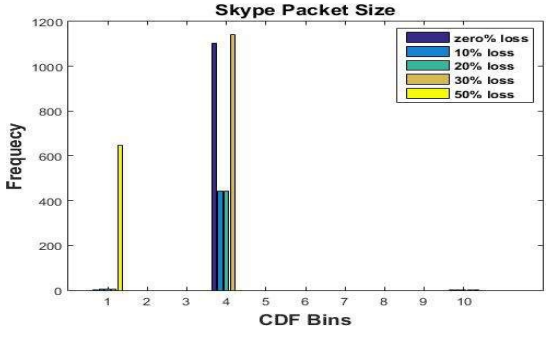


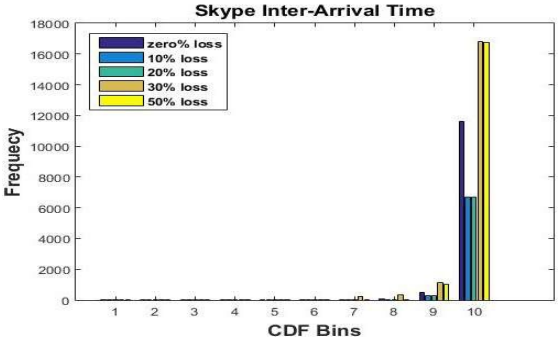




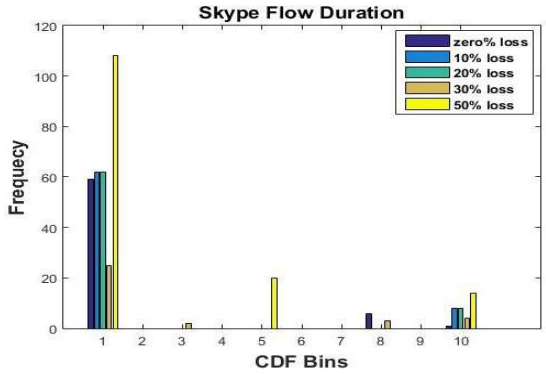


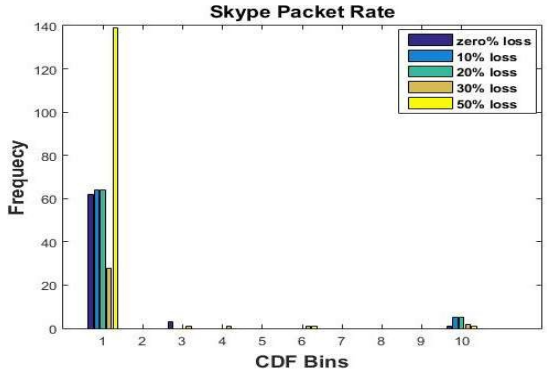


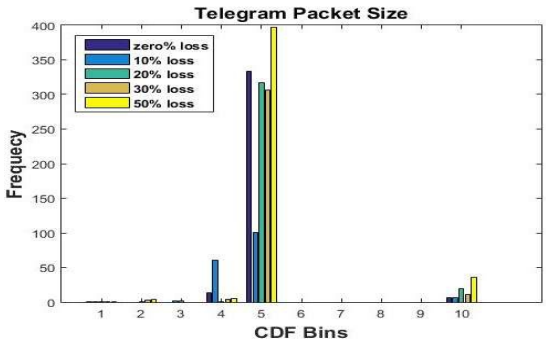


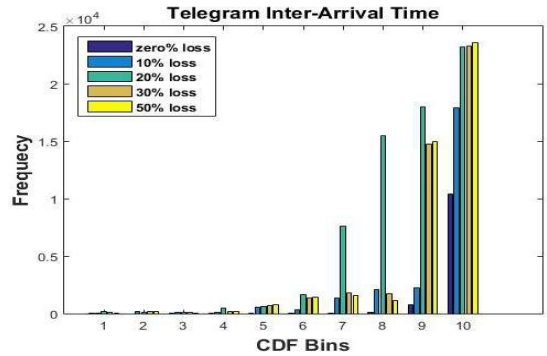


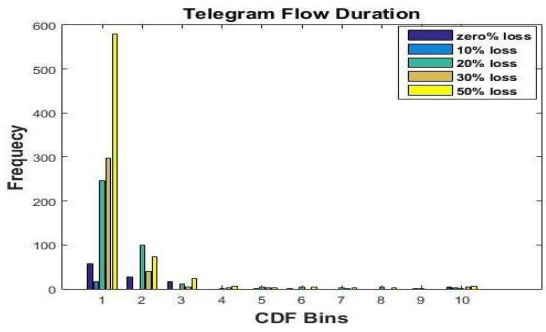
[DOI: 10.29252/jsdp.16.4.3]

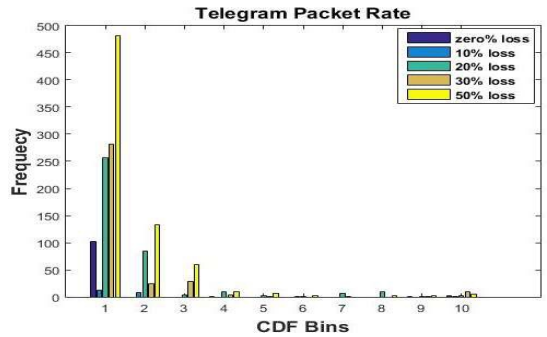


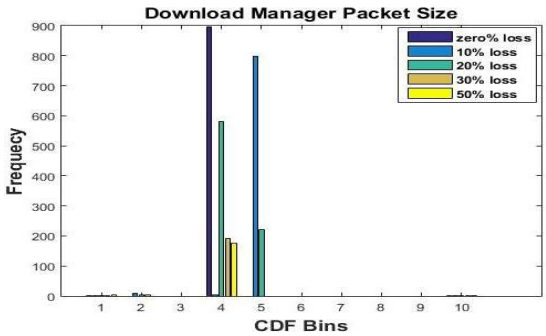


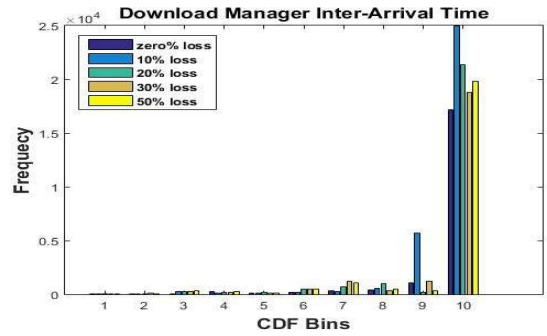


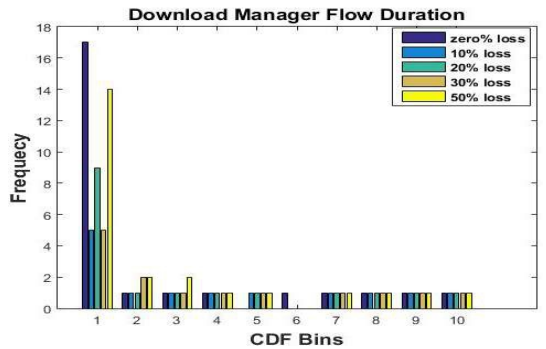


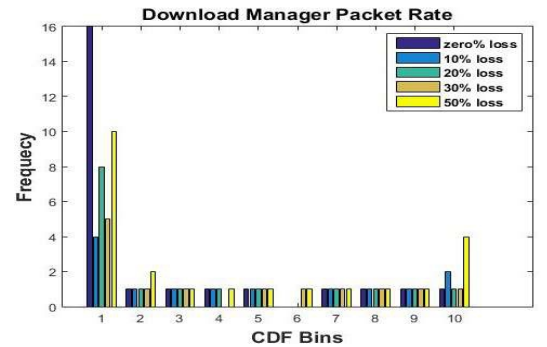




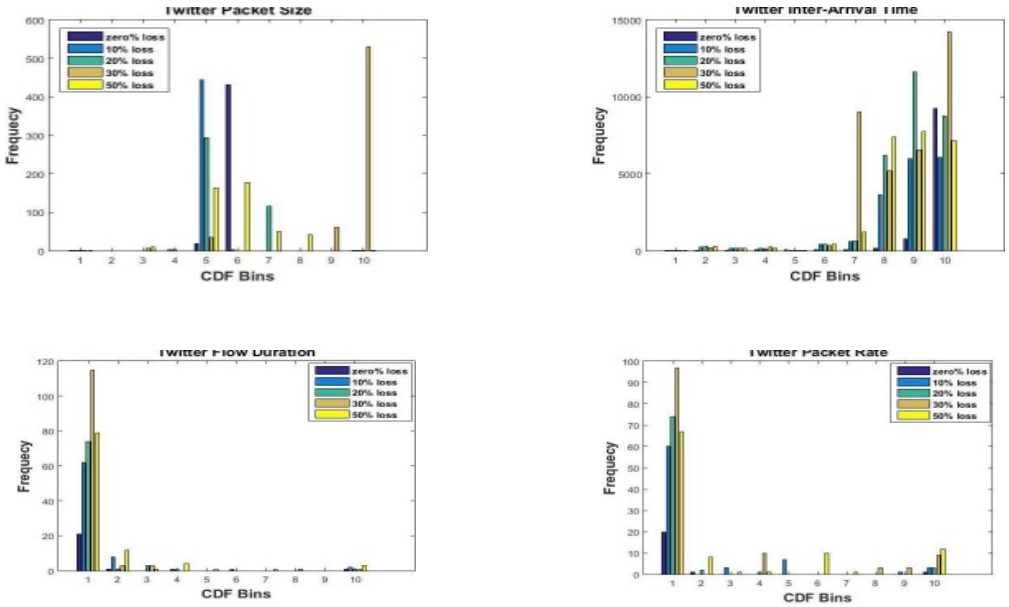








[DOI: 10.29252/jsdp.16.4.3]



(شکل-۴): فراوانی تابع توزیع تجمعی چهار ویژگی مربوط به بسته‌های هفت برنامه کاربردی در ده بازه یکسان از صفر تا یک
(Figure-4): Ferquency of Four Statistical Features CDF for Seven Applications Packet in ten equal intervals

اگر بخواهیم روش پیشنهادی را به‌صورت خلاصه بیان کنیم، مراحل زیر جهت استخراج ویژگی لازم است:

- ویژگی‌های آماری ترتیب بسته‌ها، اندازه بسته‌ها، فاصله زمانی بین بسته‌های ارسالی و دریافتی، مدت‌زمان جریان‌ها و نرخ ارسال بسته‌ها، با توجه به رفتار برنامه‌های کاربردی استخراج می‌شود [13].
- پس از استخراج ویژگی‌های آماری برای هر برنامه کاربردی، بسته‌های هر جریان برنامه کاربردی به‌عنوان یک دسته در نظر گرفته شده و برای هر دسته بردار مقادیر تابع توزیع تجمعی (CDF) محاسبه می‌شود.
- فراوانی CDFها در بازه‌های کوتاه محاسبه می‌شود.
- بدین‌ترتیب برای ترافیک شبکه هر برنامه کاربردی با احتساب تعداد نمونه‌های ورودی، تعدادی بردار ویژگی خواهیم داشت که هر کدام از این بردارها حاوی ده ویژگی (فراوانی تابع توزیع) است.
- از شبکه عصبی جهت دسته‌بندی ترافیک شبکه استفاده می‌شود.

جزئیات بیشتر درخصوص شبکه عصبی و همچنین نتایج حاصله در ادامه بعد شرح داده شده است.

۵- ارزیابی نتایج

به‌منظور ارزیابی قابلیت ویژگی‌های استخراج‌شده در شناسایی ترافیک شبکه در شرایط اتلاف بسته، آنها را در یک شبکه عصبی چندلایه پرسپترونی به‌کار می‌گیریم. مزیت اصلی استفاده از شبکه عصبی MLP قابلیت بالای این شبکه در یادگیری و پایداری آن در مقابل تغییرات ورودی است. مشخصات و همچنین تعداد لایه‌ها و ورودی‌های شبکه عصبی مورد استفاده در جدول (۲) آمده است.

(جدول-۲): مشخصات شبکه عصبی مورد استفاده
(Table-2): Characteristics of used Neural Network

تعداد نرون‌های ورودی	۱۰- تعداد ویژگی‌ها
تعداد نمونه‌های آموزش برای هر رده	۳۸۰۰
تعداد نمونه‌های آزمایش برای هر رده	۱۲۰۰
تعداد لایه های پنهان	۱
تعداد نرون در لایه پنهان	۷
نرخ یادگیری	۰/۰۵
حداکثر تعداد دفعات تکرار آموزش	۱۰۰
حد آستانه خطا برای توقف یادگیری	۰/۱

شبکه عصبی بالا پس از آموزش بر روی مجموعه داده جدول (۱) (با احتساب نرخ اتلاف برای برنامه‌های P2P تا ۳۰٪ و برای برنامه‌های Client-Server تا ۵۰٪) اجرا شده و نتایج آن در جدول (۳) نشان داده شده است. معیار ROC[1] میزان حساسیت یا پیش‌بینی درست در مقابل پیش‌بینی نادرست در یک سامانه طبقه‌بندی را نشان می‌دهد و منظور از TP[2] میزان

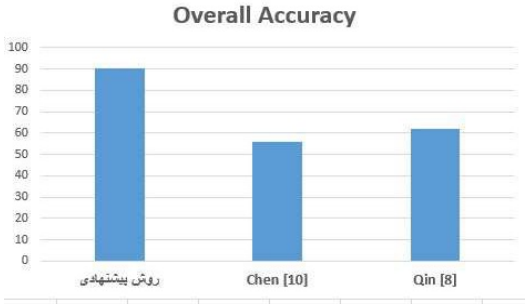
[1] Receiver Operating Characteristic
[2] True Positive

[DOI: 10.29252/jsdp.16.4.3]

تشخیص‌های درست و FP[1] میزان تشخیص‌های نادرست است.

همان‌طور که مشاهده می‌شود، با استفاده از ویژگی‌های استخراج‌شده دقت شناسایی در مجموعه‌دادهٔ حاوی اتلاف بسته به‌طور میانگین تا ۵۳ ۹۰٪ افزایش یافته و به دقت شناسایی در محیط ایده‌آل (۹۷/۵٪) نزدیک شده است. اگر بخواهیم به‌طور دقیق‌تر نمونه‌های شناسایی‌شده را مورد بررسی قرار دهیم، ماتریس درهم ریختگی[2] نتایج به‌دست‌آمده در جدول (۴) نشان داده شده است.

در جدول (۴) برای هر برنامهٔ کاربردی تعداد نمونهٔ تشخیص‌داده‌شده از هر برنامهٔ کاربردی نشان داده شده است. همان‌طورکه مشاهده می‌شود، با استفاده از ویژگی‌های استخراج‌شده دقت شناسایی نسبت به نتایج اولیه (شکل (۲)) به میزان قابل توجهی افزایش یافته و در مواردی که سامانه شناسایی نمونه مربوط به یک برنامه کاربردی را به‌اشتباه یک برنامه کاربردی دیگر تشخیص داده است، آن برنامه از خانواده همان برنامه کاربردی (خانواده برنامه‌های Client-Server یا خانواده برنامه‌های P2P) است. به‌عنوان نمونه هنگام تشخیص برنامه کاربردی Chrome اکثر نمونه‌های اشتباه تشخیص داده شده مربوط به برنامه‌های کاربردی Firefox و IE است که هر دو از خانواده برنامه‌های Client-Server بوده و رفتاری مشابه دارند، و به همین صورت در شناسایی برنامه کاربردی Skype اکثر نمونه‌های اشتباه تشخیص‌داده‌شده مربوط به برنامه‌های Telegram و IDM هستند که از خانواده برنامه‌های P2P بوده و رفتاری مشابه دارند. در ادامه کار جهت مقایسه روش پیشنهادی، میزان دقت شناسایی با استفاده از این روش با دو روش ارائه‌شده توسط Qin [8] و Chen [10] مورد ارزیابی قرار گرفت. شکل (۵) نتایج حاصل از دقت شناسایی روش پیشنهادی در مقایسه با دو روش دیگر را نشان می‌دهد.



(شکل-۵) مقایسه دقت شناسایی روش پیشنهادی با روش‌های Qin و Chen

(Figure-5): Comparison of the accuracy of proposed method with Qin and Chen methods

۶- نتیجه‌گیری و کارهای آینده

شناسایی ترافیک شبکه به‌عنوان امری ضروری در مدیریت شبکه‌های رایانه‌ای است. در شناسایی ترافیک شبکه مبتنی بر ویژگی‌های آماری، عوامل بسیاری می‌توانند بر این ویژگی‌ها تأثیر بگذارند. یکی از اصلی‌ترین مسائل تأثیرگذار، رخداد اتلاف بسته در شبکه است که به‌عنوان یکی از مباحث اصلی در کیفیت خدمات‌دهی در شبکه‌های امروزی مطرح است. در این مقاله ضمن بررسی تأثیرات اتلاف بسته در شناسایی ترافیک شبکه، ویژگی‌های آماری جدیدی از بسته‌ها استخراج شده است.

(جدول-۳): نتایج شناسایی هفت برنامه کاربردی

(Table-3): Identification Results of Seven Applications

Accuracy	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	برنامه کاربردی
74.96%	0.750	0.032	0.796	0.750	0.772	0.958	Chrome
76.5%	0.765	0.025	0.835	0.765	0.798	0.972	Firefox
94.63%	0.946	0.013	0.926	0.946	0.936	0.997	Internet Explorer
89.93%	0.899	0.014	0.912	0.899	0.906	0.983	Skype
99.39%	0.994	0.008	0.952	0.994	0.972	1.00	Telegram
98.29%	0.994	0.014	0.923	0.983	0.952	0.999	IDM
100%	1.000	0.004	0.978	1.000	0.989	0.999	Twitter
90.53%	0.906	0.015	0.907	0.886	0.903	0.986	میانگین

(جدول-۴): ماتریس درهم‌ریختگی شناسایی هفت برنامه کاربردی به کمک روش پیشنهادی

(Table-4): Confusion Matrix to Identifying seven applications using the proposed method

Twitter	IDM	Telegram	Skype	Internet Explorer	Firefox	Chrome	
2	127	79	188	172	386	2856	Chrome
0	76	40	46	91	2889	634	Firefox
0	0	6	49	3621	111	39	Internet Explorer
85	111	66	3414	27	55	38	Skype
0	0	3768	17	1	5	0	Telegram
0	3742	0	28	0	15	22	IDM
3791	0	0	0	0	0	0	Twitter

[1] False Psitive

[2] Confusion Matrix

نتایج نشان می‌دهد که رخداد اتلاف بسته می‌تواند تأثیرات چشم‌گیری بر دقت شناسایی ترافیک گذاشته و روش‌های موجود را دچار مشکل کند. ازاین‌رو با بهره‌گیری از فراوانی تابع توزیع تجمعی مربوط به رفتار داده‌های ترافیکی برنامه‌های کاربردی، ویژگی‌هایی استخراج شد. پس از ارزیابی ویژگی‌های استخراج‌شده با استفاده از شبکه عصبی بر روی مجموعه‌داده تهیه‌شده، نشان داده شد که دقت شناسایی، بهبود یافته و میزان آن به دقت شناسایی ترافیک شبکه در حالت ایده آل نزدیک شده است. به‌عنوان کارهای آینده در این زمینه می‌توان به مواردی همچون: بررسی سایر برنامه‌های کاربردی، بررسی تأثیر رخدادهای دیگر در شبکه بر روی دقت شناسایی مانند تأثیرات رمزنگاری، اثرات دیوار آتش و درنظرگرفتن حالت برخط در شناسایی ترافیک اشاره کرد.

7- References — ۷- مراجع

[1] M. Crotti, M. Dusi, F. Gringoli, L. Salgarelli, " Traffic classification through simple statistical fingerprinting", *ACM SIGCOMM Comput. Commun*, Rev. 37, pp.5–16, 2007.

[2] M. Jain, D.S. Tomar, S.K. Singh, "A Survey on TCP Congestion Control Schemes in Guided Media and Unguided Media Communication", *Int. J. Comput. Appl.* Pp.*118*, 2015.

[3] M.A. Kafi, D. Djenouri, J. Ben-Othman, N. Badache, "Congestion control protocols in wireless sensor networks: a survey", *IEEE Commun. Surv. Tutorials*, vol.16, pp.1369 1390 2014.

[4] B. Yamansavascilar, M.A. Guvensan, A.G. Yavuz, M.E. Karsligil, "Application identification via network traffic classification", *Computing, Networking and Communications (ICNC), International Conference on*, pp. 843 848, 2017.

[5] J. Kim, J. Hwang, K. Kim, K, "High-performance internet traffic classification using a Markov model and Kullback-Leibler divergence", Mob. Inf. Syst. 2016.

[6] J. Muehlstein, Y. Zion, M. Bahumi, I. Kirshenboim, R. Dubin, A. Dvir, O. Pele, "Analyzing HTTPS Encrypted Traffic to Identify User Operating System, Browser and Application" arXiv Prepr. arXiv1603.04865, 2016.

[7] H.R. Loo, S.B. Joseph, M.N. Marsono, " Online incremental learning for high bandwidth network traffic classification", *Appl. Comput. Intell. Soft Comput*, vol. 1, 2016.

[8] T. Qin, L.Wang, Z. Liu, X. Guan, "Robust application identification methods for P2P and VoIP traffic classification in backbone networks", *Knowledge-Based Syst.* Vol.82, pp.152–162, 2016.

[9] M.S. Aliakbarian, A. Fanian, F.S. Saleh, T.A. Gulliver, " Optimal supervised feature extraction in internet traffic classification" Communications, *Computers and Signal Processing (PACRIM), IEEE Pacific Rim Conference on*, pp. 102 107, 2013.

[10] Z. Chen, L. Peng, C. Gao, B. Yang, Y. Chen, J. Li, "Flexible neural trees based early stage identification for IP traffic", *Soft Comput.* Vol. 21, pp. 2035 2046, 2017.

[11] F. Ertam, E. Avci, "A new approach for internet traffic classification: GA-WK-ELM", *Measurement*, Vol. 95, pp.135–142,2017.

[12] A. Este, F. Gringoli, L. Salgarelli, "On the stability of the information carried by traffic flow features at the packet level", ACM SIGCOMM Comput. Commun. Rev. 39, 2009.

[13] M. Gandomi, H. Hassanpour, "Behavioral Analysis of Traffic Flow for an Effective Network Traffic Identification", *International Journal of Engineering (IJE), TRANSACTIONS B: Applications* ,Vol. 30, No. 11, 2017, pp.150-160.

[14] J. Bolot, "End-to-end packet delay and loss behavior in the Internet", *ACM SIGCOMM Computer Communication Review*, pp. 289–298, 1993.

[15] Z. Chen, Z. Liu, L. Peng, L. Wang, L. Zhang, "A novel semi-supervised learning method for Internet application identification", *Soft Computing*, Vol. 21 , pp. 1963 1975, 2017.

[16] N. Saqib, Y. Shakeel, M. Khan, H. Mehmood, M. Zia, "An effective empirical approach to VoIP traffic classification", *Turkish Journal of Electrical Engineering & Computer Sciences*, Vol. 25, pp. 888 900,2017.

[17] H. Shi, H. Li, D. Zhang, C. Cheng, W. Wu, "Efficient and robust feature extraction and selection for traffic classification", *Computer Networks*, Vol. 119, 2017, pp. 1—16.

[18] J. Yang, J. Deng, S. Li, Y. Hao, "Improved traffic detection with support vector machine based on restricted Boltzmann machine", *Soft Computing*, Vol. 21, pp. 3101—3112, 2017.

[DOI: 10.29252/jsdp.16.4.3]

محمدرضا گندمی متولد سال ۱۳۶۵ است. ایشان دوره کارشناسی، کارشناسی ارشد خود را به‌ترتیب در سال‌های ۱۳۸۸، ۱۳۹۱ از دانشگاه‌های آزاد اسلامی، صنعتی امیرکبیر در رشتهٔ مهندسی کامپیوتر کسب کرده است. همچنین ایشان در سال ۱۳۹۷ موفق به اخذ مدرک دکترا از دانشگاه صنعتی شاهرود شد و از جمله زمینه پژوهشی و علاقه‌مندی ایشان می‌توان به داده‌کاوی، تحلیل داده و شبکه‌های رایانه‌ای اشاره کرد.

نشانی رایانامه ایشان عبارت است از:

Ga.mohamadreza@gmail.com

حمید حسن‌پور استاد تمام دانشکده مهندسی کامپیوتر دانشگاه شاهرود هستند. ایشان در سال ۱۳۷۲ مدرک کارشناسی خود را از دانشگاه علم و صنعت و در سال ۱۳۷۵ مدرک کارشناسی ارشد خود را در گرایش هوش ماشین از دانشگاه صنعتی امیرکبیر دریافت کرد. در سال ۱۳۸۲ موفق به اخذ مدرک دکترای خود از دانشگاه صنعتی کوئینزلند استرالیا در گرایش پردازش سیگنال شد. از سال ۱۳۸۴ الی ۱۳۸۶ نامبرده به‌عنوان هیئت علمی در دانشکده مهندسی برق و کامپیوتر دانشگاه صنعتی بابل فعالیت داشت؛ سپس به دانشکده صنعتی شاهرود انتقال یافت. زمینه‌های علمی مورد علاقه ایشان پردازش سیگنال، پردازش تصویر، داده‌کاوی و پردازش متن است.

نشانی رایانامه ایشان عبارت است از:

h.hassanpour@shahroodut.ac.ir

[DOI: 10.29252/jsdp.16.4.3]