

Computer Networks traffic classification model based on DBScan clustering and gamma classification

Seyede Zohreh Majidian¹, Shiva TaghipourEivazi², Bahman Arasteh³, Ali Ghaffari⁴

¹ PhD Student. Department of Computer Engineering, Aras international Branch, Islamic Azad University, Tabriz, Iran. Email: std.z.majidian@iaut.ac.ir

² Assistant Professor. Department of Computer Engineering, Ta.C.Islamic Azad university, Tabriz, Iran

(Correspondence: taghipour@iaut.ac.ir)

³ Associate Professor. Department of Computer Engineering, Ta.C.Islamic Azad university, Tabriz, Iran.

Email: b.arasteh@iaut.ac.ir

⁴ Professor. Department of Computer Engineering, Ta.C.Islamic Azad university, Tabriz, Iran.

Email: a.ghaffari@iaut.ac.ir

ARTICLE INFO

Article history:

Article Type: Research paper

Receive: 08 June 2025

Revise: 09 August 2025

Accept: 09 September 2025

Available online: 03 October 2025

Keywords:

Traffic classification
machine learning
DBScan clustering
gamma classification

ABSTRACT

Traffic classification is a crucial network monitoring process with wide applications in security, quality of service, and network management. With the increasing complexity and variety of network traffic, new challenges arise, including the lack of labeled training data. To address this challenge, this paper presents a traffic classification mechanism that combines unsupervised and semi-supervised machine learning algorithms. This mechanism uses a limited set of labeled training data to improve classification accuracy. The proposed method represents each traffic flow as a feature vector containing the statistical characteristics of that flow. The number of features generated for each sample is reduced using principal component analysis. DBScan clustering is employed to determine the correct traffic type for each untagged traffic stream. Finally, the gamma classifier model is used to separate the new traffic flows. The efficiency of the proposed method has been evaluated using real data sets. The results show that the proposed method can classify traffic flows with an average accuracy of 95.12%, representing at least a 7.03% improvement over previous approaches.

Cite this article: Majidian, S. Z., TaghipourEivazi, S., Arasteh, B., & Ghaffari, A. (2025). Computer NetworksTraffic Classification Model Based on DBScan Clustering and Gamma Classification. *Electronic and Cyber Defense*, 13(3), 1-16. 2025.

DOI: <https://dor.isc.ac/dor/20.1001.1.23224347.1404.13.3.1.0>



OPEN ACCESS

© The Author(s). retain the copyright and full publishing rights

Publisher: Imam Hossein University

مدل طبقه‌بندی ترافیک شبکه‌های کامپیوتری مبتنی بر خوشه‌بندی DBScan و طبقه‌بند گاما

سیده زهره مجیدیان¹، شیوا تقی پور عیوضی²، بهمن آراسته³، علی غفاری⁴

¹ دانشجوی دکتری، گروه مهندسی کامپیوتر، واحد بین‌المللی ارس، دانشگاه آزاد اسلامی، تبریز، ایران. std.z.majidian@iaut.ac.ir

² استادیار، گروه مهندسی کامپیوتر، واحد تبریز، دانشگاه آزاد اسلامی، تبریز، ایران. (نویسنده مسئول: taghipour@iau.ac.ir)

³ دانشیار، گروه مهندسی کامپیوتر، واحد تبریز، دانشگاه آزاد اسلامی، تبریز، ایران. b.arasteh@iaut.ac.ir

⁴ استاد، گروه مهندسی کامپیوتر، واحد تبریز، دانشگاه آزاد اسلامی، تبریز، ایران. a.ghaffari@iaut.ac.ir

چکیده	مشخصات مقاله
طبقه‌بندی ترافیک یکی از مهم‌ترین فرآیندهای نظارت بر شبکه است که کاربردهای گسترده‌ای در حوزه‌های امنیت، کیفیت خدمات و مدیریت شبکه دارد. با افزایش پیچیدگی و تنوع ترافیک شبکه، چالش‌های جدیدی از جمله کمبود داده‌های آموزشی برچسب به وجود می‌آید. به منظور رفع این چالش در این مقاله، سازوکار طبقه‌بندی ترافیک با ترکیب الگوریتم‌های یادگیری ماشین بدون نظارت و نیمه نظارتی ارائه می‌شود. این سازوکار از مجموعه محدودی از داده‌های آموزشی برچسب‌گذاری شده برای بهبود دقت طبقه‌بندی استفاده می‌کند. روش پیشنهادی، هر جریان ترافیک را به عنوان یک بردار ویژگی توصیف می‌کند که شامل ویژگی‌های آماری آن جریان است. تعداد ویژگی‌های ایجاد شده برای هر نمونه نیز با استفاده از تحلیل مؤلفه‌های اصلی کاهش می‌یابد. خوشه‌بندی DBScan برای تعیین نوع ترافیک صحیح برای هر جریان ترافیک بدون برچسب استفاده می‌شود. در نهایت، از مدل طبقه‌بندی گاما برای تفکیک جریان‌های ترافیک جدید استفاده می‌شود. کارایی روش پیشنهادی با استفاده از مجموعه داده‌های واقعی ارزیابی شده است. نتایج نشان می‌دهد که روش پیشنهادی قادر به طبقه‌بندی جریان‌های ترافیکی با دقت متوسط 95.12 درصد است که حداقل 7.03 درصد بهبود را نسبت به رویکردهای قبلی نشان می‌دهد.	تاریخچه مقاله: دریافت: 1404/03/18 بازنگری: 1404/05/13 پذیرش: 1404/06/16 ارائه آنلاین: 1404/07/16
	کلیدواژه‌ها: طبقه‌بندی ترافیک یادگیری ماشین خوشه‌بندی DBScan طبقه‌بندی گاما

استناد: مجیدیان، سیده زهره، تقی پور عیوضی، شیوا، آراسته، بهمن، و غفاری، علی. (۱۴۰۴). مدل طبقه‌بندی ترافیک شبکه‌های کامپیوتری مبتنی بر خوشه‌بندی DBScan و طبقه‌بند گاما. پدافند الکترونیکی و سایبری. ۱۳(۳)، ص ۱۶-۱. <https://dor.isc.ac/dor/20.1001.1.23224347.1404.13.3.1.0>

۱. مقدمه

خوشه‌بندی DBScan برای کشف الگوهای پنهان در داده‌های بدون برچسب، روشی جدید و کارآمد برای شناسایی انواع مختلف ترافیک ارائه می‌دهد که در تحقیقات پیشین این حوزه مورد توجه قرار نگرفته است. همچنین، طبقه‌بندی گاما به‌عنوان یک مدل یادگیری با نظارت، توانایی بالایی در تشخیص تفاوت‌های ظریف بین انواع مختلف ترافیک را دارد. این ترکیب منحصر به فرد، به‌ویژه در شرایطی که داده‌های برچسب‌گذاری شده محدود است، عملکرد مطلوبی از خود نشان می‌دهد و می‌تواند چالش‌های ناشی از شباهت زیاد بین برخی از انواع ترافیک را برطرف نماید. در مجموع، این روش، یک رویکرد قوی و انعطاف‌پذیر برای طبقه‌بندی جریان‌های ترافیک شبکه است که می‌تواند در سناریوهای دنیای واقعی مورد استفاده قرار گیرد.

ادامه این مقاله به شرح زیر سازمان‌دهی شده است: بخش 2 شامل بررسی تحقیقات مرتبط است. جزئیات روش پیشنهادی در بخش 3 توضیح داده شده است. بخش چهارم به نتایج و بحث‌ها اختصاص دارد. بخش 5 شامل نتیجه‌گیری و ارائه طرحی کلی برای کارهای آینده است.

۲. مروری بر رویکردهای پیشین

تجزیه و تحلیل ترافیک ترکیبی از روش‌هایی است که به دنبال تعیین پیکربندی‌های نادرست، ناهنجاری‌ها، الگوها و روابط در داده‌های ترافیکی شبکه است [7] و هدف آن تعیین نوع ترافیک یا نام برنامه کاربردی است. این راهبردها می‌توانند ترافیک را در کلاس‌های از پیش تعریف‌شده مانند غیرعادی یا عادی دسته‌بندی کنند. دقت روش‌ها برای طبقه‌بندی ترافیک به نحوه ردیابی ترافیک بستگی دارد که این امر خود به نمونه‌های زمینه، یعنی برنامه‌ها یا پروتکل‌ها با برچسب‌های صحیح متکی است. جمع‌آوری داده‌های معتبر از جریان ترافیک ضروری است زیرا اساس آموزش و ارزیابی عملکرد است [8]. امروزه طبقه‌بندی ترافیک به یک مسئله چالش‌برانگیز تبدیل شده است و دلیل آن ظهور روش‌های جدید مانند رمزنگاری و کپسوله‌سازی است که عملکرد راهبردهای طبقه‌بندی مرسوم را کاهش می‌دهد [9]. محققان در [10] بر انتخاب ویژگی‌های بهینه در مسئله طبقه‌بندی ترافیک رمزگذاری شده تمرکز کرده‌اند و رویکردی هدفمند برای این منظور ارائه کرده‌اند. در این روش، رویکردهای انتخاب ویژگی با جزئیات برای مجموعه داده‌های مختلف تجزیه و تحلیل شده است. در روش ارائه‌شده در مقاله [11] مدل ماشین بردار پشتیبان حساس به هزینه برای حل این مشکل معرفی شده است. رویکرد ارائه‌شده ترکیبی از ماشین بردار پشتیبان چند کلاسه و روش یادگیری فعال است که از وزن دهی پویا برای بهبود عملکرد ماشین بردار پشتیبان استفاده می‌کند. دقت طبقه‌بندی و عملکرد الگوریتم ارائه‌شده با استفاده از دو مجموعه داده MOORE_SET و NOC_SET بررسی شده است.

طبقه‌بندی ترافیک شبکه کاربردهای زیادی از جمله کنترل ترافیک، امنیت شبکه و تحلیل شبکه دارد [1]. یکی از مشکلات مهم در زمینه طبقه‌بندی ترافیک شبکه، کمبود داده‌های برچسب‌گذاری شده آموزشی است. برای آموزش مدل‌های طبقه‌بندی ترافیک شبکه، نیاز به داده‌های زیادی است که هر کدام با برچسب صحیح مشخص شده باشند. جمع‌آوری چنین داده‌هایی بسیار دشوار و زمان‌بر است. راه‌حل‌های مختلفی برای حل این مشکل ارائه شده است. استفاده از روش‌های یادگیری تقویتی [2] یکی از راه‌حل‌ها است که در این روش مدل به‌صورت آزمایش و خطا یاد می‌گیرد. این روش می‌تواند برای آموزش مدل‌های طبقه‌بندی ترافیک شبکه بدون نیاز به داده‌های برچسب‌گذاری شده استفاده شود. روش دیگر، به‌کارگیری روش‌های یادگیری بدون نظارت یا نیمه نظارتی است. در یادگیری بدون نظارت (نیمه نظارتی)، الگو (مدل) با استفاده از داده‌های بدون برچسب (یا نمونه‌های محدود دارای برچسب) یادگیری را انجام می‌دهد. این روش می‌تواند برای شناسایی الگوهای موجود در داده‌ها استفاده شود [3]. از طرفی، لزوم بررسی شماره درگاه، داده‌های کاربر یا رفتار میزبان در شبکه از جمله محدودیت‌هایی هستند که روش‌های پیشین برای طبقه‌بندی ترافیک شبکه با آن مواجه هستند و دسترسی به این اطلاعات در شبکه‌های عمومی منجر به نقض حریم خصوصی کاربران می‌شود. هدف این پژوهش، ارائه روشی جدید برای طبقه‌بندی جریان‌های ترافیک شبکه با استفاده از روش‌های یادگیری ماشین و بدون نیاز به بررسی شماره درگاه، داده‌های کاربر یا رفتار میزبان است. این الگوریتم، از ترکیبی از روش‌های یادگیری ماشین [4]، [5] بدون نظارت و نظارت‌شده استفاده می‌کند. در مرحله اول، از تحلیل مؤلفه‌های اصلی برای کاهش ابعاد ویژگی‌های آماری جریان استفاده می‌شود. این کار، حجم داده‌های مورد نیاز برای آموزش مدل را کاهش می‌دهد و آن را برای سناریوهای بلادرنگ قابل اجرا می‌کند. در مرحله دوم، از خوشه‌بندی بدون نظارت توسط DBScan¹ [6] برای تخمین برچسب نمونه‌های آموزشی بدون برچسب استفاده می‌شود. این کار، دقت طبقه‌بندی را در صورت کمبود داده‌های برچسب‌گذاری بهبود می‌بخشد. در نهایت، طبقه‌بندی جریان توسط یک مدل طبقه‌بندی گاما انجام می‌شود. این مدل، دقت طبقه‌بندی را در داده‌های ترافیک واقعی بهبود می‌بخشد. رویکرد پیشنهادشده برای طبقه‌بندی ترافیک شبکه در این مقاله، نوآوری‌های متعددی را ارائه می‌دهد که به‌طور مؤثری چالش‌های دنیای واقعی در این حوزه را برطرف می‌نماید. ترکیب نوآورانه ارائه‌شده از راهبردهای یادگیری بدون نظارت و با نظارت، این امکان بهره‌مندی از حداکثر پتانسیل داده‌های برچسب‌گذاری شده و بدون برچسب را می‌دهد. به‌ویژه، استفاده از الگوریتم

¹ Density-Based Spatial Clustering of Applications with Noise

ادغام‌شده‌اند و طبق نتایج حاصل‌شده، شبکه‌های بیزین می‌تواند با دقت متوسط 96.79 درصد از سایر طبقه‌بندی‌کننده‌ها بهتر عمل کند. رویکرد مورد‌استفاده در این مقاله به‌کارگیری یادگیری عمیق به‌منظور طبقه‌بندی ترافیک بر اساس ویژگی‌های استخراج‌شده خودکار است که با استفاده از طبقه‌بندی Distiller سامانه قادر به حل مشکل ناهمگونی و غلبه بر محدودیت‌های یادگیری چندوجهی می‌شود. در مقابل مزایای این رویکرد، می‌توان به ترافیک چندوجهی (رمزگذاری شده) اشاره نمود که منجر به حمل اطلاعاتی در شبکه می‌شود که باعث اختلال در شبکه خواهد شد. طبقه‌بندی ترافیک رمزگذاری شده دستگاه‌های تلفن همراه در [18] ارائه‌شده است. این رویکرد شامل چهار مرحله قطعه‌بندی جریان، استخراج ویژگی‌ها، فیلتر کردن ویژگی‌ها و طبقه‌بندی ویژگی‌ها است که با استفاده از شبکه عصبی کانولوشن عمیق انجام می‌شود. در ادامه در مقاله [19] نیز مدلی برای طبقه‌بندی ترافیک رمزگذاری شده توسط یادگیری عمیق چندوظیفه‌ای چندوجهی پیشنهادشده است که می‌تواند ارتباط بین داده‌های ورودی و کلاس‌های هدف را بدون تقسیم مسئله به مشکلات فرعی مانند انتخاب یا استخراج ویژگی‌ها، مدل‌سازی کند. مزایای این رویکرد ارزیابی استحکام طبقه‌بندی‌کننده ترافیک شبکه، مبتنی بر یادگیری عمیق در برابر ترافیک متخاصم است. از طرفی وجود ترافیک متخاصم در طبقه‌بندی‌کننده‌های ترافیک شبکه مبتنی بر یادگیری عمیق با استفاده از روش‌های تولید آشفته‌گی متخاصم جهانی، منجر به پیش‌بینی اشتباه می‌شود. روش‌های یادگیری عمیق همچنین می‌توانند سازوکار پیش‌پردازش را خودکار کنند [20]. تحقیق انجام‌شده در [21]، کاربرد روش‌های مبتنی بر یادگیری عمیق را برای طبقه‌بندی ترافیک بررسی کرده است.

۳. ارائه روش پیشنهادی

در این بخش مراحل روش پیشنهادی شرح داده‌شده است. در رویکرد پیشنهادی، ترکیب روش‌های بدون نظارت و نیمه نظارتی به‌منظور طبقه‌بندی ترافیک استفاده‌شده است. برای این هدف، مجموعه محدودی از جریان‌های ترافیکی برچسب‌گذاری شده برای بهبود عملکرد طبقه‌بندی استفاده می‌شود.

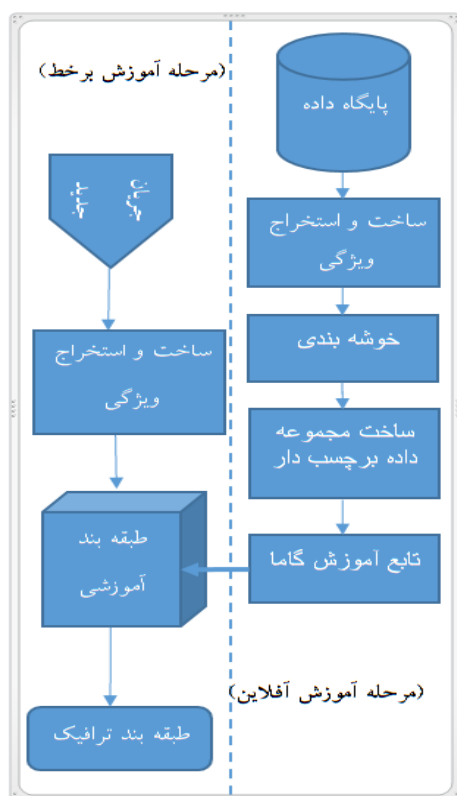
رویکرد پیشنهادی شامل پنج مرحله اصلی است:

- 1- توصیف ویژگی‌های جریان‌های ترافیکی بر چسب‌دار و بدون برچسب
- 2- استخراج ویژگی‌ها توسط تجزیه و تحلیل مؤلفه‌های اصلی^۱
- 3- خوشه‌بندی بدون نظارت بردارهای ویژگی با استفاده از DBScan

نتایج بیانگر این است که الگوریتم ارائه‌شده منجر به کاهش هزینه محاسبات و نیز افزایش دقت طبقه‌بندی‌شده است. روش ارائه‌شده در مقاله [12] باهدف طبقه‌بندی ترافیک رمزگذاری شده، از یک شبکه عصبی عمیق با ساختار موازی در شبکه تقویت‌شده است و پس از هر لایه کانولوشن یک شبکه میکرو را برای بهبود مدل‌سازی محلی اتخاذ می‌کند. این طبقه‌بندی‌کننده مدل یادگیری عمیق مبتنی بر معماری شبکه عصبی کانولوشن است. هر جزء کانولوشن به‌عنوان یک شبکه میکرو برای بهبود مدل‌سازی محلی عمل می‌کند. همچنین این مدل لایه‌های کاملاً متصل را با لایه‌های ادغام میانگین سرا سری جایگزین کرده است که در ساده‌سازی مدل مؤثر است و تعداد پارامترهای مدل را به میزان قابل‌توجهی کاهش می‌دهد. نتایج آزمایش‌های روی مجموعه داده‌های ترافیک رمزگذاری شده ISCX VPN-nonVPN بیانگر این است که مدل پیشنهادی تعادل بهتری بین دقت طبقه‌بندی و پیچیدگی مدل را داراست. نویسندگان در مقاله [13]، از روش‌های یادگیری گروهی برای طبقه‌بندی ترافیک شبکه ایمن استفاده کرده‌اند. دو مدل مجموعه‌ای تقویت‌گرایان و درخت تصمیم‌گیری برای طبقه‌بندی ترافیک در اتصالات استفاده‌شده است. ترکیبی از ویژگی‌های طیفی و زمانی ترافیک شبکه برای طبقه‌بندی در [14] استفاده‌شده است. در این روش، شبکه‌های رمزگذار خودکار مکرر با شبکه عصبی کانولوشن بازگشتی ترکیب می‌شوند تا رابطه بین ویژگی‌های مختلف شبکه و ویژگی‌های طیفی را توصیف کنند. همچنین دنباله این ویژگی‌ها به‌عنوان ویژگی‌های زمانی در نظر گرفته می‌شود که باهم ترکیب‌شده و برای طبقه‌بندی استفاده می‌شوند. روش ارائه‌شده که از روی هم قرار دادن رمزگذار خودکار با یک شبکه عصبی کاملاً متصل شکل گرفته است، به بهبود ۲۸ درصدی در معیار دقت نسبت به رویکردهای مطالعه شده قبلی مبتنی بر یادگیری ماشین دست‌یافته است. یک مطالعه تجربی برای طبقه‌بندی ترافیک شبکه اینترنت اشیا با استفاده از یادگیری ماشین در مقاله [15] انجام‌شده است. این تحقیق از مجموعه داده‌ای از ترافیک شبکه دستگاه‌های اینترنت اشیا در یک بازه زمانی 20 روزه استفاده کرده و دقت الگوریتم‌های مختلف یادگیری ماشین را در طبقه‌بندی این داده‌ها گزارش کرده است. در [16] از سازوکار یادگیری انتقال عمیق برای طبقه‌بندی ترافیک در شبکه‌های اینترنت اشیا نسل پنجم استفاده‌شده است. این رویکرد، طبقه‌بندی را با تنظیم دقیق شبکه عصبی و سازوکار انتقال وزنی پیکربندی می‌کند. بر اساس این راه‌حل، کل داده‌های آموزشی موردنیاز نخواهد بود و تنها با استفاده از 10 درصد از داده‌ها، می‌توان به دقت طبقه‌بندی مشابه دست‌یافت. در رویکرد ارائه‌شده در مقاله [17] روش‌های مختلف یادگیری ماشین از جمله ماشین بردار پشتیبان و شبکه‌های بیزین ساده در شبکه‌های تعریف‌شده نرم‌افزاری برای طبقه‌بندی ترافیک داده‌ها

¹ Principal Component Analysis (PCA)

این طبقه بند گاما برای طبقه‌بندی جریان‌های ترافیکی جدید (نمونه‌های آزمایشی) استفاده خواهد شد. بدین ترتیب، در مرحله نظارت‌شده روش پیشنهادی، بردار ویژگی نمونه‌های آزمایشی ساخته‌شده و ویژگی‌ها استخراج می‌شوند (مطابق بخش الف). سپس، ویژگی‌های استخراج‌شده به مدل طبقه بند گامای آموزش‌دیده ارسال می‌شود. مراحل در شکل (2) نشان داده‌شده است.



شکل (2). مراحل آموزش آفلاین و طبقه‌بندی برخط جریان ترافیک در روش پیشنهادی

الگوریتم (1) شبه کد رویکرد پیشنهادی را نشان می‌دهد. در این الگوریتم، هر مرحله از رویکرد پیشنهادی به بخش مربوطه خود در مقاله مرتبط شده است.

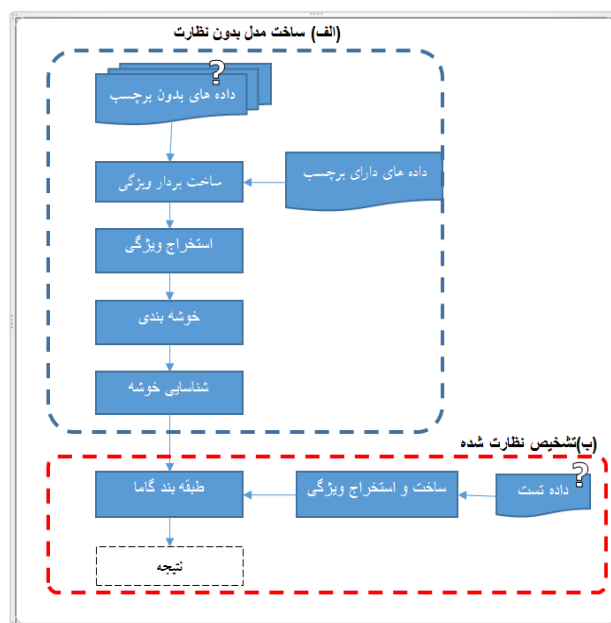
الگوریتم (1): شبه کد روش پیشنهادی

- گام 1: ایجاد بردار ویژگی برای هر نمونه ترافیک (بخش 4-2-1)
- گام 2: استخراج ویژگی‌های مناسب توسط تجزیه و تحلیل مؤلفه‌های اصلی (بخش 4-2-2)
- گام 3: اعمال خوشه‌بندی DBScan بر روی ویژگی‌های استخراج‌شده (بخش 4-2-3)
- گام 4: اعمال مراحل زیر برای هر کلاستر (بخش 4-2-4)
 - 1-4: شمارش تعداد ترافیک‌ها بر اساس نمونه‌های برچسب دار خوشه
 - 2-4: برچسب‌گذاری خوشه بر اساس بیشترین فراوانی برچسب‌های خوشه

4- شناسایی خوشه‌ها با شمارش داده‌های برچسب‌دار در هر خوشه

5- طبقه‌بندی جریان‌های ترافیک جدید توسط طبقه بند گاما

شکل (1) بیانگر مراحل روش پیشنهادی است. مطابق این شکل چارچوب پیشنهادی شامل ترکیبی از داده‌های جریان ترافیک برچسب‌دار و بدون برچسب به همراه ماژول‌هایی برای ساخت ویژگی، استخراج ویژگی، خوشه‌بندی و طبقه‌بندی است. مجموعه داده برچسب‌گذاری شده شامل مجموعه محدودی از جریان‌های ترافیک و انواع آن‌ها است.



شکل (1). مراحل رویکرد پیشنهادی

در بخش الف، هر نمونه پایگاه داده توسط یک بردار ویژگی متشکل از ویژگی‌های آماری جریان ترافیک آن توصیف می‌شود. در مرحله بعد ابعاد بردارهای ویژگی ایجادشده برای هر نمونه با استفاده از الگوریتم تجزیه و تحلیل مؤلفه‌های اصلی کاهش می‌یابد. هدف از این مرحله، بهبود سرعت پردازش نمونه‌ها در مراحل بعدی و همچنین کاهش اثر مخرب ویژگی‌های نامرتبط بر نتیجه طبقه‌بندی است. در مرحله بعد، نمونه‌های پایگاه داده توسط الگوریتم DBScan خوشه‌بندی می‌شوند. بنابراین، جریان‌های ترافیکی با ویژگی‌های آماری مشابه در خوشه‌های مشابه گروه‌بندی می‌شوند. مرحله بعدی شناسایی هر خوشه خواهد بود، به این معنی که نوع ترافیک صحیح برای هر خوشه تعیین می‌شود. برای انجام این کار، نمونه‌های برچسب‌گذاری شده در هر خوشه شمارش می‌شود و به هر خوشه برچسبی بر اساس اکثریت اختصاص می‌یابد. پس‌ازاین مرحله، داده‌های بدون برچسب، برچسب خوشه خود را خواهند گرفت.

مجموعه داده‌های برچسب‌گذاری شده برای ساخت طبقه‌بندی گاما برای طبقه‌بندی نظارت‌شده (بخش ب) استفاده خواهد شد.

جدول (1). ویژگی‌های آماری بکار رفته در روش پیشنهادی

ویژگی	تعداد ویژگی	نوع ویژگی	توصیف
بسته‌های نمونه‌برداری شده	$2 \times k$	زمانی	اندازه بسته‌های منتقل شده تا نقطه زمانی فعلی در هر جهت جریان
بایتهای نمونه‌برداری شده	$2 \times k$	زمانی	بایتهای به نقطه زمانی فعلی در هر جهت جریان منتقل می‌شوند
اندازه نمونه‌برداری	$4 \times k$	زمانی	میانگین و انحراف معیار اندازه بسته تا نقطه زمانی فعلی در هر جهت جریان
زمان نمونه‌برداری	$4 \times k$	زمانی	میانگین و انحراف معیار زمان بین بسته تا نقطه زمانی فعلی در هر جهت جریان
بسته‌های جریان	2	جریان	تعداد بسته‌های منتقل شده در جهت جریان تا زمان فعلی
بایتهای جریان	2	جریان	کل بایتهای منتقل شده در هر جهت جریان تا زمان فعلی
اندازه جریان	6	جریان	میانگین، حداقل و حداکثر اندازه بسته در هر جهت جریان تا زمان فعلی
جریان زمانی	6	جریان	میانگین، حداقل و حداکثر زمان بین بسته‌ها در هر جهت جریان تا زمان جاری
طول بردار ویژگی			$16 + 10k$

۳-۲. استخراج ویژگی با استفاده از تجزیه و تحلیل

مؤلفه‌های اصلی

برخی از ویژگی‌های ساخته شده ممکن است به نوع ترافیک شبکه مرتبط نباشد. این ویژگی‌ها علاوه بر کاهش سرعت پردازش نمونه، باعث کاهش دقت طبقه‌بندی نیز خواهد شد. بنابراین، در مرحله دوم رویکرد پیشنهادی از تجزیه و تحلیل مؤلفه‌های اصلی برای استخراج ویژگی‌ها و کاهش اثر مخرب ویژگی‌های نامربوط به انواع ترافیک استفاده می‌شود.

با در نظر داشتن ماتریس داده X^T با مقدار میانگین تجربی صفر، که هر سطر آن نشان‌دهنده یک نمونه و هر ستون آن مشخص‌کننده یک ویژگی است، تحلیل مؤلفه‌های اصلی را می‌توان به صورت زیر تعریف نمود [23]:

$$Y^T = X^T W = VS \quad (1)$$

در رابطه (1)، VSW^T نشان‌دهنده تجزیه مقدا رهای منفرد ماتریس X^T است. بر اساس تعریف اولیه تحلیل مؤلفه‌های اصلی، هدف از این الگوریتم تبدیل ماتریس داده X با ابعاد M به ماتریس داده Y با ابعاد L است. بنابراین فرض می‌شود که ماتریس X از بردارهای ستونی $X_1, \dots, X_N$ تشکیل شده است که هریک از این بردارهای ستونی معرف یک ویژگی در X است. بدین ترتیب، ماتریس داده‌های X دارای ابعاد $M \times N$ خواهد بود. با در نظر

3-4: اختصاص برچسب خوشه به هر یک از اعضای بدون برچسب خوشه

گام 5: آموزش طبقه‌بندی کننده گاما را با استفاده از نمونه‌های آموزشی پایگاه داده (بخش 4-2-5)

گام 6: طبقه‌بندی نمونه‌های آزمایشی با استفاده از مدل طبقه‌بندی کننده گامای آموزش دیده

۳-۱. ساخت بردار ویژگی

مرحله اول در رویکرد پیشنهادی، توصیف هر جریان ترافیکی توسط یک بردار ویژگی با طول ثابت است. در روش پیشنهادی، یک مجموعه داده برچسب‌دار کوچکی از جریان‌های ترافیک را در کنار یک مجموعه بدون برچسب بزرگ در نظر گرفته شده است. تفاوت بین این دو مجموعه داده در این است که مجموعه داده برچسب‌گذاری شده دارای یک فیلد اضافی در مورد کلاس‌های ترافیکی است. برچسب‌های این نمونه‌ها به صورت دستی تنظیم شده است و هر نمونه دارای یکی از برچسب‌های 1- مرور، 2- ایمیل، 3- چت، 4- استریم، 5- انتقال فایل، 6- VoIP، و 7- P2P است [22].

برای تعریف بردار ویژگی هر نمونه، نمونه‌های بدون برچسب و برچسب‌دار باهم ترکیب می‌شوند تا یک مجموعه داده بدون برچسب جدید را تشکیل دهند. سپس ویژگی‌های آماری این نمونه‌ها در هر دو جهت جریان، یعنی روبه‌جلو (کلاینت به سرور) و عقب (سرور به کلاینت) استخراج می‌شود. در روش پیشنهادی، دو گروه آماری از ویژگی‌ها برای توصیف یک جریان ترافیک استفاده می‌شود:

1- ویژگی‌های استخراج شده از نقاط زمانی خاص در جریان‌ها. این ویژگی‌های آماری می‌توانند ویژگی‌های مبتنی بر نمونه یک جریان ترافیک را تعریف کنند و از نقاط زمانی مشخص شده به طور تصادفی در طول اتصال استخراج می‌شوند. در این ویژگی‌ها، نقطه زمانی برای استخراج ویژگی k به صورت یک عدد تصادفی با توزیع نمایی و میانگین 2^k (ثانیه) تعیین می‌شود. به عنوان مثال، اگر k روی 10 تنظیم شود، 10 نقطه تصادفی با توزیع نمایی و مقادیر میانگین 2^1 تا 2^{10} ثانیه انتخاب می‌شود. سپس ویژگی‌های آماری این نقاط تصادفی از جریان‌ها استخراج می‌شود. در ادامه این گروه ویژگی‌های زمانی نامیده می‌شوند.

2- ویژگی‌هایی که با توجه به ویژگی‌های کلی جریان محاسبه می‌شوند. این ویژگی‌های آماری که ویژگی‌های کلی یک جریان را توصیف می‌کنند، می‌توانند در تشخیص نوع ترافیک شبکه نیز مفید باشند. شایان ذکر است که این نوع ویژگی‌ها نیازی به اتمام جریان ندارند. این ویژگی‌ها در واقع ویژگی‌های آماری تا زمان فعلی هستند (و برخلاف ویژگی‌های زمانی، به نقاط زمانی خاصی بستگی ندارند). در ادامه این گروه ویژگی‌های زمانی نامیده می‌شوند. مجموعه ویژگی‌های مورد استفاده برای ساخت بردار ویژگی‌ها در روش پیشنهادی در جدول (1) توضیح داده شده است.

انتخاب مقدار l باید به صورتی باشد که دارای مقدار کمینه بوده و در عین حال g از مقدار قابل قبولی برخوردار باشد. مثلاً می‌توان حداقل l را به صورتی انتخاب نمود که $g[m=l] \leq 90\%$.

هـ) انتقال داده به فضای مختصات جدید: بدین منظور باید ابتدا تبدیلات زیر را اعمال نمود:

- ماتریس $S_{M,l}$ که نشان‌دهنده انحراف معیار مجموعه داده است، به صورت $S[i] = \sqrt{C[i,i]}$ محاسبه می‌گردد.
- داده به صورت کسر $Z = \frac{B}{S}$ تبدیل می‌شود.
- داده‌ها بر اساس رابطه زیر به فضای جدید نگاشت می‌یابند [23]:

$$Y = W^* \cdot ZY = W^* \cdot Z \quad (7)$$

پس از استخراج ویژگی‌های بارز بر اساس روند فوق، در گام بعدی از راهبرد خوشه‌بندی به‌منظور مقابله با مسئله محدودیت داده‌های دارای برچسب استفاده می‌شود.

۳-۳- خوشه‌بندی بدون نظارت توسط DBScan

در این مرحله از رویکرد پیشنهادی، از روش *DBScan* برای خوشه‌بندی بدون نظارت ویژگی‌های استخراج‌شده استفاده می‌شود که نمونه‌های پرت از داده‌های معتبر تفکیک‌شده و بدین ترتیب می‌توان نمونه‌های پرت را در فرآیند برچسب‌گذاری نمونه‌ها فیلتر نمود. الگوریتم *DBScan* برای خوشه‌بندی نمونه‌ها بر اساس هر شاخص از دو پارامتر آستانه همسایگی ϵ و آستانه تعداد نقاط P بهره می‌گیرد و شامل گام‌های زیر است [24]: **گام ۱:** مجموعه‌ی خوشه‌های C را به‌صورت تهی مقداردهی کن.

گام ۲: برای هر نمونه مشاهده نشده x از مجموعه داده ورودی گام‌های زیر را تکرار کن:

گام ۳: داده x را به مجموعه داده‌های مشاهده‌شده اضافه کن و نمونه‌های همسایه x را بر اساس آستانه همسایگی ϵ در N ذخیره کن.

گام ۴: اگر $|N| < P$ ، آنگاه x را به‌عنوان نویز نشانه‌گذاری کن و به گام ۲ برو؛ در غیر این صورت گام بعدی را اجرا کن.

گام ۵: در صورتی که $|N| \geq P$ ، آنگاه x را به مجموعه C اضافه کن. سپس، برای هر نمونه مانند $x' \in N$ گام‌های زیر را اجرا کن:

گام ۶: نمونه x' را از N حذف کن. سپس در صورتی که x' مشاهده‌شده است، به گام ۹ برو و در غیر این صورت گام‌های زیر را تکرار کن:

گام ۷: داده x' را به مجموعه مشاهده‌شده اضافه کن و نمونه‌های همسایه x' را بر اساس آستانه همسایگی ϵ در N' ذخیره کن.

گام ۸: در صورتی که $|N| \geq P$ ، آنگاه دو مجموعه N و N' را به‌صورت مجموعه N ادغام کن ($N = N \cup N'$).

گام ۹: اگر برای x' خوشه‌ای تعیین نشده است، آن را در خوشه-ی C قرار بده و به گام ۵ برو

داشتن این ساختار، مراحل استخراج مؤلفه‌های اصلی X بر اساس ماتریس کوواریانس به‌صورت گام‌های زیر خواهد بود:

الف) محاسبه میانگین تجربی داده و نرمال‌سازی آن: میانگین تجربی داده، برداری است که به‌صورت رابطه زیر محاسبه می‌شود [23]:

$$u[m] = \frac{1}{N} \sum_{i=1}^N X[m,i] \quad (2)$$

میانگین تجربی داده در رابطه بالا، به‌صورت مشخص بر روی سطرهاى ماتریس اعمال می‌شود. پس از آن ماتریس فاصله داده با میانگین تجربی به‌صورت زیر محاسبه می‌شود [23]:

$$B = X - uh \quad (3)$$

در رابطه بالا، h یک بردار تماماً یک با اندازه $1 \times N$ است.

ب) محاسبه ماتریس کوواریانس: در گام دوم، ماتریس کوواریانس C دارای $M \times M$ از طریق رابطه زیر محاسبه می‌شود [23]:

$$C = E[B \otimes B] = E[B \cdot B^*] = \frac{1}{N} B \cdot B^* \quad (4)$$

در رابطه فوق، E مشخص‌کننده میانگین حسابی است. همچنین عملگر $\otimes$ نشان‌دهنده ضرب خارجی ماتریس بوده و B^* ترانپوز مزدوج ماتریس B است.

ج) تخمین مقادیر ویژه-ماتریس کوواریانس و جایگشت بردارهای ویژه: در این گام، مقادیر ویژه و بردارهای ویژه در ماتریس کوواریانس C ، با استفاده از رابطه زیر محاسبه می‌شوند [23]:

$$V^{-1}CV = D \quad (5)$$

در رابطه فوق، V ماتریس بردارهای ویژه بوده و D یک ماتریس قطری است که درایه‌های واقع در قطر اصلی آن مقادیر ویژه-ماتریس هستند. هر مقدار ویژه در این ماتریس با یک بردار ویژه متناظر است. به عبارتی ماتریس V دارای ابعاد $M \times M$ بوده و ستون‌های آن بردارهای ویژه هستند؛ به صورتی که بردار ویژه V_q در ستون q ام این ماتریس قرار دارد و مقدار ویژه q ام ماتریس D متناظر با آن است ($\lambda_q = D_{q,q}$). جایگشت بردارهای ویژه بر اساس بزرگی مقادیر ویژه متناظر با آن‌ها صورت می‌پذیرد. در این حالت بر اساس ترتیب نزولی مقادیر ویژه، بردارهای ویژه مرتب‌سازی می‌شوند.

د) انتخاب یک زیرمجموعه از بردارهای ویژه به‌عنوان مجموعه پایه: انتخاب یک زیرمجموعه از بردارهای ویژه از طریق تحلیل مقادیر ویژه صورت می‌پذیرد. زیرمجموعه نهایی بر اساس جایگشت حاصل از مرحله قبل به‌صورت $V_1, \dots, V_l$ تعیین می‌شود. در این گام می‌توان از انرژی تجمعی استفاده نمود که بر اساس آن [23]:

$$g[m] = \sum_{q=1}^m \lambda_q \quad (6)$$

تعریف (2) اعمال عملگر آلفا بر بردارها: با فرض اینکه $x, y \in A^n$ بردارهای ستونی ورودی باشند. خروجی $\alpha(x, y)$ بردار n بعدی است که مؤلفه‌های آن به شرح زیر محاسبه می‌شود:

$$(8) \alpha(x, y)_i = \alpha(x_i, y_i) \quad \alpha(x, y)_i = \alpha(x_i, y_i)$$

تعریف (3) عملگر u_β : با فرض اینکه الگوی باینری $x \in A^n$ به‌عنوان ورودی، این عملگر عدد صحیح غیر منفی زیر را به‌عنوان خروجی تولید کرده و به‌صورت زیر محاسبه می‌شود:

$$(9) u_\beta(x) = \sum_{i=1}^n \beta(x_i, x_i)$$

تعریف (4) عملگر هرس: با فرض اینکه $x \in A^n$ و $y \in A^m$ و $m < n$ ، دو بردار باینری باشند. سپس y هرس شده توسط x ، به‌صورت $\mathcal{M}_x$ نمایش داده‌شده و برداری دودویی و n بعدی است که مؤلفه‌های آن به‌صورت زیر محاسبه شده است.

$$(10) (y \parallel x)_i = y_{i+m-n}, \quad (i = 1, 2, \dots, n)$$

عملگر گاما به یک بردار باینری به‌عنوان ورودی نیاز دارد. برای مقابله با بردارهای حقیقی یا عدد صحیح، روشی برای نمایش این بردارها به‌صورت باینری موردنیاز است. در این کار، معمولاً از کد جانسون-موبیوس اصلاح‌شده در [25] استفاده می‌شود. با توجه به اینکه جزئیات کامل این الگوریتم در [25] موردبحث قرار گرفته است، لذا از پرداختن به روند آن در این صرف‌نظر می‌شود.

تعریف (5) عملگر گاما: عملگر شباهت گاما دو الگوی باینری مانند $x \in A^n$ و $y \in A^m$ و $n \leq m$ و یک عدد صحیح غیرمنفی مانند θ را به‌عنوان ورودی گرفته و به ازای هریک از دو حالت زیر، یک خروجی دودویی تولید می‌کند:

حالت 1: اگر $n = m$ ، خروجی مطابق رابطه زیر محاسبه می‌شود [26]:

$$(11) \gamma(x, y, \theta) = \begin{cases} 1, & \text{if } m - u_\beta[\alpha(x, y) \bmod 2] \leq \theta \\ 0, & \text{otherwise} \end{cases}$$

در رابطه فوق، $\bmod$ نشانگر عملکرد باقیمانده تقسیم است.

حالت 2: اگر $n < m$ ، بجای y با استفاده از $y \parallel x$ به‌صورت رابطه زیر محاسبه می‌شود [26]:

$$(12) \gamma(x, y, \theta) = \begin{cases} 1, & \text{if } m - u_\beta[\alpha(x, y \parallel x) \bmod 2] \leq \theta \\ 0, & \text{otherwise} \end{cases}$$

با در نظر داشتن عملگرهای فوق، در ادامه به تشریح عملکرد طبقه بند گاما پرداخته می‌شود. با فرض اینکه الگوهای پایه به‌صورت $\{(x^u, y^u) | u = 1, 2, \dots, p\}$ دارای کاردینالیت p بوده و الگوی آزمون به‌صورت $\tilde{x} \in R^n$ توصیف شده باشد. در این حالت، طبقه بند گاما از طریق گام‌های محاسباتی زیر نمونه آزمایشی $\tilde{x}$ را طبقه‌بندی می‌کند:

پس از دسته‌بندی نمونه‌ها، تمامی نمونه‌های نويز نادیده گرفته شده و برای سایر نمونه‌های خوشه‌بندی‌شده، عمل شناسایی دسته و انتشار برچسب صورت می‌پذیرد که در ادامه تشریح شده است.

۳-۴. شناسایی خوشه‌ها

پس از خوشه‌بندی نمونه‌های مجموعه داده ورودی، برچسب هر خوشه حاصل باید بر اساس نوع ترافیک تعیین شود. برای انجام این کار، از نتایج خوشه‌بندی نمونه‌های برچسب‌گذاری شده استفاده خواهد شد. به این ترتیب توزیع نمونه‌های برچسب‌دار در نتیجه خوشه‌بندی بررسی شده و بیشترین تعداد برچسب در هر خوشه شمارش می‌شود. بنابراین، برچسب خوشه، برچسبی است که بیشترین فراوانی نمونه‌های برچسب‌دار را در آن دارد. اگر بیش از یک برچسب بیشترین فراوانی را در خوشه داشته باشد، یک برچسب به‌طور تصادفی از بین آن‌ها انتخاب می‌شود. پس از انتخاب برچسب Li برای یک خوشه Ci ، تمام نمونه‌های بدون برچسب در Ci به‌عنوان Li برچسب‌گذاری می‌شوند. نتیجه این مرحله یک مجموعه داده کامل دارای برچسب خواهد بود. این مجموعه داده برای طبقه‌بندی نیمه نظارتی در مرحله بعدی روش پیشنهادی استفاده خواهد شد.

۳-۵. طبقه‌بندی جریان‌های ترافیکی توسط طبقه بند گاما

در آخرین گام از روش پیشنهادی، از طبقه بند گاما به‌منظور ساخت مدل طبقه‌بندی جریان‌های ترافیکی استفاده خواهد شد. این مدل طبقه‌بندی، یک روش نظارت‌شده است که نام آن از عملگر شباهتی که استفاده می‌کند، عملگر گاما گرفته شده است. تعمیم‌یافته این عملگر دو بردار دودویی x و y و یک عدد صحیح مثبت مانند θ را به‌عنوان ورودی دریافت کرده و اگر هر دو بردار مشابه باشد مقدار 1 و در غیر این صورت مقدار 0 را برمی‌گرداند. عملگر گاما از عملگرهای دیگری (مانند α ، β و ...) استفاده می‌کند که در ادامه به تعریف آن‌ها پرداخته می‌شود.

تعریف (1) عملگرهای آلفا و بتا: با توجه به مجموعه $A = \{0, 1\}$ و $B = \{0, 1, 3\}$ ، عملگرهای آلفا (α) و بتا (β) به‌صورت جدول (2) خواهند بود.

جدول (2). عملگرهای آلفا و بتا در طبقه بند گاما

$\alpha: Ax \rightarrow B$		
x	y	$\alpha(x, y)$
0	0	1
0	1	0
1	0	2
1	1	1
$\beta: Bx \rightarrow A$		
x	y	$(x, y)\beta$
0	0	0
0	1	0
1	0	0
1	1	1
2	0	1
2	1	1

گام ۱۱) در غیر این صورت الگوی $\tilde{x}$ به کلاس دارای اولین مقدار بیشینه اختصاص خواهد یافت.

۴. مقایسه و ارزیابی

در این مطالعه، از نمونه‌های مجموعه داده ISCXVPN2016 [27] برای ارزیابی عملکرد روش پیشنهادی استفاده شد و شامل مجموعه‌ای از داده‌های ترافیکی واقعی است که از نظر کمیت و تنوع غنی است. نمونه‌های این مجموعه داده بر اساس سناریوی آلیس و باب ثبت شده است. برای انجام این کار، دو حساب کاربری (آلیس و باب) ایجاد شده و جریان ترافیک این کاربران در استفاده از سرویس‌هایی مانند اسکایپ، پوتیوب، Hangout و غیره ضبط شده است. برنامه Wireshark برای ضبط جریان‌های ترافیک از دستگاه‌های شبکه استفاده شده است و هر داده گرفته شده به‌عنوان یک فایل pcap یا pcapng در مجموعه داده ذخیره می‌شود. در ادامه انواع ترافیک و برنامه‌های کاربردی مورد استفاده در مجموعه داده ISCXVPN2016 برای هر نوع ترافیک تشریح خواهد شد.

ISCXVPN2016 حاوی داده‌های ترافیکی ضبط شده برای 7 کلاس ترافیک در شبکه‌های VPN و اتصالات غیر VPN است. بنابراین، تعداد کل کلاس‌های منحصربه‌فرد در این مجموعه داده برابر با 14 خواهد بود: Email, VPN-Email, Chat, VPN-Chat و غیره. در این مطالعه، از 4739 جریان از مجموعه داده‌های ISCXVPN2016 استفاده شده است که 60 جریان از هر کلاس برچسب‌گذاری شده است. بنابراین تعداد کل جریان‌های برچسب‌گذاری شده 420 است. لیست انواع ترافیک این مجموعه داده عبارت‌اند از:

1. Browsing
2. Email
3. Chat
4. Streaming
5. File Transfer
6. VoIP
7. P2P

برای ثبت این نوع ترافیک‌ها از Wireshark و tcpdump استفاده شده است و 150 فایل pcap/pcapng با حجم کل 28 گیگابایت ضبط شده است.

روش پیشنهادی در محیط MALTAB 2020a پیاده‌سازی شده است. از tshark (بخشی از برنامه Wireshark) برای ساخت بردارهای ویژگی فایل‌های مجموعه استفاده شده است. با توجه به تعداد محدود پارامترهای مدل پیشنهادی، به‌منظور پیکربندی این پارامترها از راهبرد جستجوی جامع و بررسی تجربی نتایج استفاده شده است. در این روند، ترکیبات مختلف مقادیر پارامترها مورد بررسی قرار گرفته و پیکربندی که منجر به بیشینه نمودن عملکرد مدل گردد، انتخاب شده است. این مجموعه پارامترها عبارت‌اند از: مقدار پارامتر k در گام توصیف ویژگی (که مشخص‌کننده طول بردار ویژگی‌های مبتنی بر زمان

گام ۱) مجموعه پایه آموزشی را با استفاده از کد جانسون-موبیوس اصلاح شده به فرم دودویی تبدیل نموده تا برای هر مؤلفه در بردارهای n -بعدی مجموعه پایه یک مقدار e_m به‌صورت زیر محاسبه شود [26]:

$$e_m(j) = \max_{i=1}^p(x_j^i), \quad \forall j \in \{1, 2, \dots, n\} \quad (13)$$

گام ۲) پارامتر توقف به‌صورت $\rho = \max_{j=1}^n[e_m(j)]$ محاسبه می‌شود.

گام ۳) الگوی آزمون نیز با استفاده از کد جانسون-موبیوس اصلاح شده و با همان پارامترهایی که برای کدگذاری مجموعه اصلی استفاده شده است، کدگذاری خواهد شد. اگر هر e_j به‌دست آمده از $e_m(j)$ متناظرش بیشتر باشد، آن را با استفاده از تعداد بیشتری بیت، کد می‌شود.

گام ۴) شاخص همه الگوهای پایه را به دو شاخص تبدیل می‌شود: یکی برای کلاس آن‌ها و دیگری برای موقعیت آن‌ها در کلاس.

گام ۵) مقدار اولیه پارامتر θ برابر صفر تنظیم می‌شود.

گام ۶) اگر $\theta = 0$ ، آنگاه با محاسبه $\gamma(x_j^u, \tilde{x}_j, 0)$ بررسی می‌شود که آیا $\tilde{x}$ یک الگوی پایه است و سپس وزن انباشته اولیه c_u^0 برای هر الگوی پایه به‌صورت زیر محاسبه می‌شود [26]:

$$c_u^0 = \sum_{i=1}^n \gamma(x_j^u, \tilde{x}_j, 0), \quad \text{for } u = 1, 2, \dots, p \quad (14)$$

اگر یک مقدار بیشینه یکتا برابر با n وجود دارد؛ آنگاه کلاس متناظر با این مقدار بیشینه به الگوی آزمون اختصاص داده می‌شود [26]:

$$\tilde{y} = y^\sigma \quad \text{such that} \quad \max_{i=1}^p c_i^0 = c_\sigma^0 = n \quad (15)$$

گام ۷) مقدار $\gamma(x_j^{\omega}, \tilde{x}_j, \theta)$ را برای هر مؤلفه الگوهای پایه محاسبه می‌شود.

گام ۸) مقدار مجموع وزن دار c_i را برای هر کلاس با استفاده از رابطه زیر محاسبه خواهد شد [26]:

$$c_i = \frac{\sum_{\omega=1}^{k_i} \sum_{j=1}^n \gamma(x_j^{\omega}, \tilde{x}_j, \theta)}{k_i} \quad (16)$$

در رابطه فوق، k_i نشان‌دهنده کاردینالیت مجموعه پایه کلاس i است.

گام ۹) اگر بیش از یک مقدار بیشینه در بین c_i های مختلف وجود داشته باشد، مقدار θ را یک واحد افزایش داده و گام‌های 7 و 8 را تا زمانی که تنها یک مقدار بیشینه وجود داشته باشد؛ یا شرط خاتمه $\theta > \rho$ برآورده گردد؛ تکرار خواهد شد.

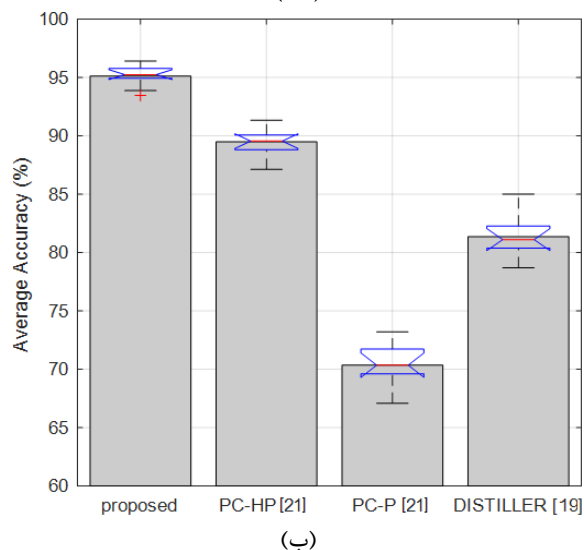
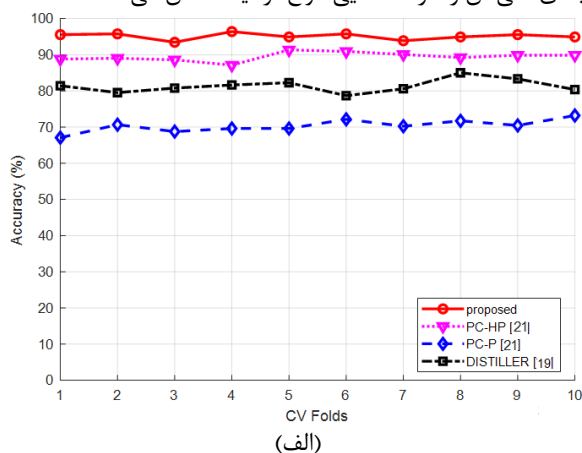
گام ۱۰) اگر یک مقدار بیشینه یکتا در بین c_i های مختلف وجود داشته باشد؛ آنگاه کلاس مربوط به الگوی آزمایشی $\tilde{x}$ با استفاده از رابطه زیر محاسبه می‌شود [26]:

$$\tilde{y} = y^j \quad \text{such that} \quad \max c_i = c_j \quad (17)$$

شکل 3. (الف) دقت طبقه‌بندی روش پیشنهادی به ازای تغییرات تعداد ویژگی و (ب) توزیع نمونه‌ها در دسته‌های هدف بر اساس سه مؤلفه برتر استخراج‌شده توسط تجزیه و تحلیل مؤلفه‌های اصلی

شکل (3). الف، نتایج به‌دست‌آمده را نشان می‌دهد. همچنین در شکل (3). ب، توزیع نمونه‌ها در دسته‌های هدف بر اساس سه مؤلفه برتر استخراج‌شده توسط الگوریتم تجزیه و تحلیل مؤلفه‌های اصلی نمایش داده شده است.

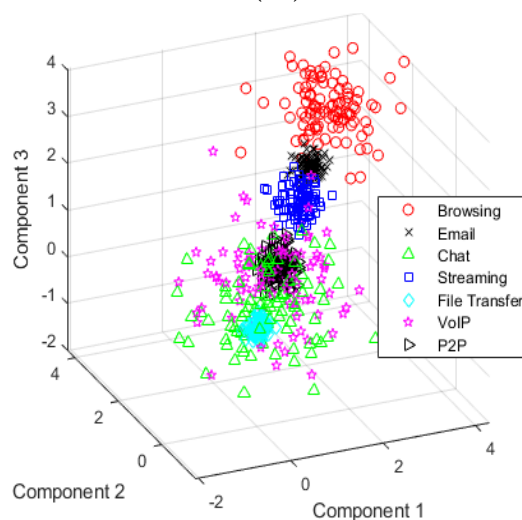
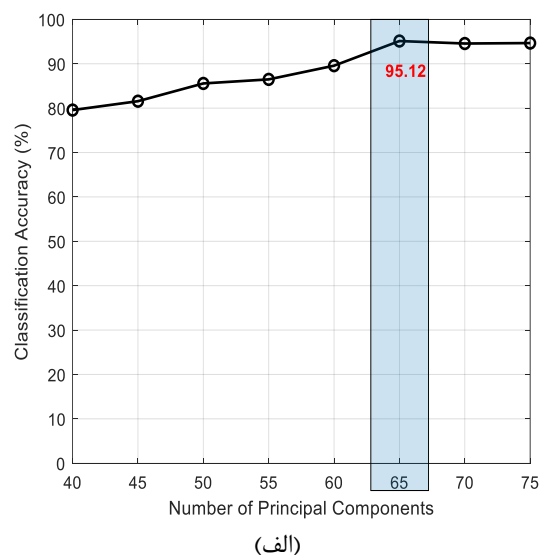
شکل (4). الف نتایج طبقه‌بندی روش پیشنهادی را از نظر دقت، در طول 10 با تکرار آزمایش‌ها نشان می‌دهد. در این شکل، عملکرد مدل پیشنهادی در طی تکرارهای مختلف با روش‌های PC-HP [21]، PC-P [21] و DISTILLER [19] مقایسه شده است. با توجه به نتایج این شکل، دقت روش پیشنهادی در طول این تکرارها، برتر از روش‌های مقایسه شده و میانگین دقت آن در طبقه‌بندی تمام نمونه‌های داده 95.12 درصد است که توانایی بالای کلی آن را در شناسایی انواع ترافیک نشان می‌دهد.



شکل (4). مقایسه دقت طبقه‌بندی (الف) تغییرات دقت در طول 10 بار تکرار آزمایش‌ها و (ب) نمودار جعبه‌ای تغییرات دقت

مطابق شکل (3). الف، حداکثر مقدار دقت زمانی به دست می‌آید که تعداد ویژگی‌های استخراج‌شده برابر با 65 باشد. طبق نتایج تجربی، استفاده از مقادیر بالاتر برای تعداد ویژگی‌های

استخراج‌شده از جریان ترافیکی است)، ابعاد ویژگی استخراج‌شده توسط الگوریتم تجزیه و تحلیل مؤلفه‌های اصلی و تعداد تکرار در روند اعتبارسنجی متقاطع. در طی فرآیند ساخت بردار ویژگی، مقدار k برابر 10 در نظر گرفته شده است که 4739 نمونه و بردار ویژگی به طول $116=10 \times 10 + 16$ برای توصیف هر یک به‌دست‌آمده است. این مقدار منجر به شکل‌گیری بردارهای ویژگی با غنای اطلاعاتی کافی برای توصیف خصوصیات مرتبط با نوع جریان‌های ترافیکی خواهد شد. از طرفی، بررسی مقادیر مختلف پارامتر تعداد تکرار (به ازای مقادیر 5، 10، 20 و 30 تکرار) در اعتبارسنجی متقاطع اختلاف معناداری در نتایج نهایی را نشان نداد. لذا آزمایش‌های این بخش بر اساس 10 بار تکرار در رویکرد اعتبارسنجی متقاطع انجام شده است. علاوه، راهبرد جستجوی جامع نشان داد که در روش پیشنهادی، کاهش ابعاد پایگاه داده به 65 ویژگی با استفاده از الگوریتم تجزیه و تحلیل مؤلفه‌های اصلی بهترین عملکرد را در پی دارد. برای این کار، دقت روش پیشنهادی برای مقادیر مختلف این پارامتر بررسی شده است.

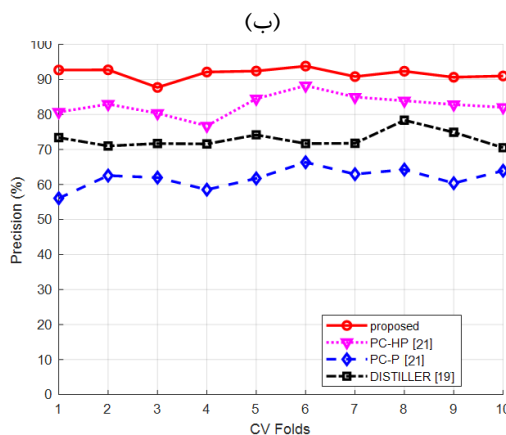
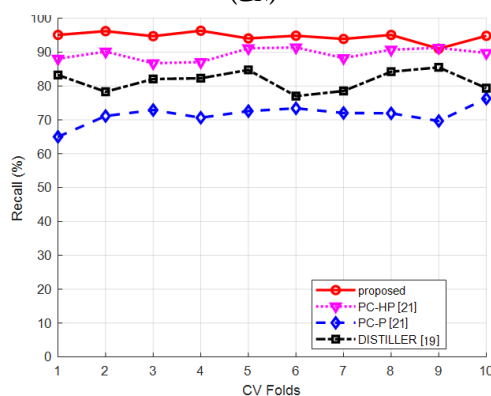
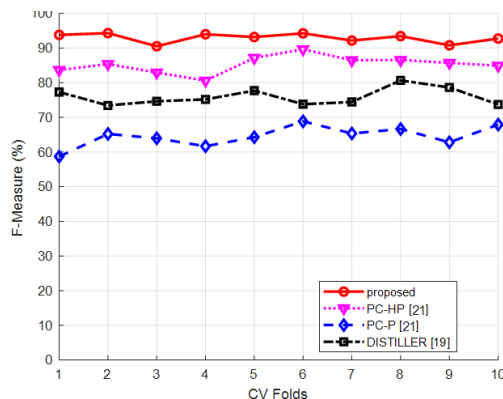


(ب)

در این معادله، FN (منفی کاذب) به نمونه‌های متعلق به کلاس هدف اشاره دارد که در کلاس‌های دیگر طبقه‌بندی شده‌اند. با استفاده از مقادیر صحت و پوشش، میانگین هماهنگ این معیارها را می‌توان به صورت F-Measure محاسبه کرد:

$$F - Measure = \frac{2 \times Precision \times Recal}{Precision + Recal} \quad (20)$$

مطابق نمودارهای ترسیم‌شده در شکل (5)، مقادیر صحت، پوشش و F-Measure در روش پیشنهادی پس از 10 بار تکرار آزمایش‌ها به ترتیب برابر با 91.07، 94.59 و 93.05 است. این مقادیر ارتقای قابل توجهی را در مقایسه با روش‌های مقایسه شده نشان می‌دهد. مقادیر بالاتر صحت روش پیشنهادی در مقایسه با سایر روش‌ها نشان می‌دهد که به صورت میانگین و برای هر دسته هدف، نسبت بالاتری از برچسب‌های پیش‌بینی شده برای آن دسته صحیح بوده است.



استخراج شده، طبقه‌بندی روش پیشنهادی را بهبود نمی‌بخشد. بنابراین، در ادامه این آزمایش‌ها، روش پیشنهادی با استفاده از این مقادیر بهینه ارزیابی خواهد شد. از طرفی، شکل (3) ب. به خوبی نشان می‌دهد که با بهره‌گیری از مؤلفه‌های اصلی استخراج شده می‌توان نمونه‌های متعلق به دسته‌های مختلف را از هم تفکیک نمود. شکل (3) الف نشان می‌دهد که با در نظر گرفتن مؤلفه‌های اصلی بیشتر این قابلیت تفکیک نیز تقویت می‌شود و براین اساس می‌توان کارایی راهبرد تجزیه و تحلیل مؤلفه‌های اصلی در توصیف کارآمد ویژگی‌ها را نتیجه گرفت.

از طرفی، در شکل (4) ب، مقادیر میانگین دقت به دست آمده توسط روش‌های مقایسه شده در کنار بازه تغییرات دقت آن‌ها با استفاده از یک نمودار جعبه‌ای نمایش داده شده است. در این نمودار، تغییرات دقت در طی تکرارهای مختلف به چارک‌های مختلف تقسیم‌بندی شده و مقدار میانه دقت به صورت نقطه میانی هر جعبه نمایش داده شده است. بررسی این نمودار نشان می‌دهد که بازه تغییرات دقت روش پیشنهادی در طی 10 تکرار اعتبارسنجی متقاطع به بازه 93.46 تا 96.41 محدود شده است که محدوده فشرده‌تری را در مقایسه با روش‌های مقایسه شده از خود نشان می‌دهد. از طرفی، مقادیر میانه و میانگین دقت به دست آمده توسط روش پیشنهادی در سطوح بالاتری نسبت به روش‌های مقایسه شده قرار دارند که تأیید کننده قابلیت اعتماد بالاتر مدل پیشنهادی در پیش‌بینی صحیح نوع جریان‌های ترافیکی شبکه است.

در شکل‌های (5) الف تا (5) ج، مقادیر میانگین F-measure، پوشش و صحت در طول تکرارهای اعتبارسنجی متقاطع نشان داده شده است. این نتایج برای روش پیشنهادی در مقایسه با روش‌های PC-P [21]، PC-HP [21] و DISTILLER [19] است.

از آنجایی که طبقه‌بندی ترافیک یک مسئله چند کلاسه است، ابتدا مقادیر دقت، پوشش و F-Measure برای هر کلاس به طور جداگانه محاسبه شده و سپس میانگین ارزش این معیارها محاسبه شده است. در هنگام محاسبه این معیارها برای هر کلاس، آن طبقه مثبت و سایر طبقات منفی در نظر گرفته شده است. در این مورد، معیار صحت می‌تواند دقت الگوریتم را برای هر کلاس به طور جداگانه توصیف کند:

$$Precision = \frac{TP}{TP + FP} \times 100 \quad (18)$$

در این معادله، TP تعداد نمونه‌های طبقه مورد مطالعه را که به درستی شناسایی شده‌اند (مثبت صحیح)، نشان می‌دهد و FP به نمونه‌های متعلق به سایر طبقات که در کلاس مثبت طبقه‌بندی شده‌اند، اشاره می‌کند (مثبت کاذب). معیار پوشش نسبت نمونه‌های مثبتی را نشان می‌دهد که الگوریتم می‌تواند به درستی شناسایی کند:

$$recall = \frac{TP}{TP + FN} \times 100 \quad (19)$$

(ج)

شکل (5). مقادیر میانگین (الف) F-Measure، (ب) پوشش و (ج) صحت، در 10 بار تکرار آزمایش‌ها در روش‌های ارزیابی شده این ویژگی تأیید می‌کند که به‌صورت کلی برچسب‌های پیش‌بینی‌شده توسط روش پیشنهادی از احتمال صحت بالاتری برخوردارند. در مقابل، بالاتر بودن مقدار پوشش روش پیشنهادی در طی تکرارهای مختلف نشان می‌دهد که این راهبرد قادر بوده است تا نرخ بالاتری از نمونه‌های متعلق به هر دسته ترافیکی را به‌درستی شناسایی کند. به بیانی دیگر، انواع جریان‌های ترافیکی با احتمال کمتری در سایر دسته‌ها طبقه‌بندی خواهند شد.

درنهایت، سطوح بالاتر F-measure کسب‌شده توسط روش پیشنهادی تأیید می‌کند که این راهکار از جنبه‌های عملکردی بررسی‌شده دارای برتری کلی نسبت به روش‌های مقایسه‌شده پیشین است. تجزیه و تحلیل دقیق‌تر عملکرد مدل پیشنهادی از طریق بررسی این معیارها به تفکیک هر دسته ممکن خواهد بود.

جدول (3) نتایج ارزیابی الگوریتم پیشنهادی برای طبقه‌بندی جریان‌های ترافیکی مجموعه داده ISCXPVN2016 را خلاصه می‌کند. در این جدول معیارهای مورد مطالعه برای هر کلاس ترافیکی به تفکیک محاسبه‌شده است. همچنین، کیفیت طبقه‌بندی رویکرد پیشنهادی با کارهای قبلی مقایسه شده است. روش‌های مقایسه‌شده نیز از همین مجموعه داده برای ارزیابی رویکردهای خود استفاده کرده‌اند.

مطابق این جدول، روش پیشنهادی بهتر از الگوریتم‌های مقایسه‌شده عمل می‌کند. بهترین مقدار دقت در روش پیشنهادی، متعلق به کلاس‌های ترافیک «Browsing»، «Chat» و «streaming» است. این بدان معنی است که خروجی‌های ساخته‌شده با روش پیشنهادی در این کلاس‌ها با احتمال بالایی درست هستند. همچنین، بهترین مقدار پوشش برای رویکرد پیشنهادی، متعلق به همین کلاس‌ها است. این بدان معنی است که روش پیشنهادی می‌تواند اغلب این نمونه‌های ترافیک را به‌درستی شناسایی کند.

جدول (3). مقایسه کیفیت طبقه‌بندی روش پیشنهادی

		Browsing	Email	Chat	Streaming
روش پیشنهادی	%Precision	98.9	89.2	94.5	96.2
	%Recall	95.8	93.2	96.7	93.9
	%F-Measure	97.3	91.16	95.59	95.04
PC-HP[21]	%Precision	-	65.77	85.35	79.96
	%Recall	-	76.45	76.45	91.50
	%F-Measure	-	80.66	80.66	85.34
PC-P [21]	%Precision	-	59.97	74.11	48.08
	%Recall	-	69.84	47.31	41.88
	%F-Measure	-	64.53	57.75	44.77
DISTILLER [19]	%Precision	-	78.28	61.12	78.96
	%Recall	-	77.17	72.13	75.78
	%F-Measure	-	77.72	66.17	77.34
		Transfer	VoIP	P2P	Average
روش پیشنهادی	%Precision	79.9	92.3	90.9	91.70
	%Recall	92.7	95.7	94.1	94.5
	%F-Measure	85.8	93.9	92.4	93.0
PC-HP[21]	%Precision	98.4	87.24	99.3	86.01
	%Recall	90.2	84.47	99.97	86.51
	%F-Measure	94.1	85.83	99.63	87.71
PC-P[21]	%Precision	67.5	76.33	69.66	65.94
	%Recall	77.1	85.15	77.33	66.45
	%F-Measure	72.0	80.50	73.29	65.47
DISTILLER [19]	%Precision	82.3	80.53	93.33	79.09
	%Recall	75.2	73.74	97.77	78.63
	%F-Measure	78.6	76.99	95.50	78.72

می‌دهد. این ماتریس نتایج طبقه‌بندی تمامی نمونه‌های مجموعه داده در تمامی تکرارهای ارزیابی متقاطع را نشان می‌دهد. در این

شکل (6) ماتریس درهم‌ریختگی به‌دست‌آمده از طبقه‌بندی جریان‌های ترافیکی شبکه توسط مدل پیشنهادی را نشان

توسط روش پیشنهادی با برچسب صحیح مقداردهی شده‌اند (پوشش بالاتر) و درعین حال؛ با مقایسه این ماتریس‌ها بر اساس ستون‌های تشکیل‌دهنده می‌توان دریافت که در روش پیشنهادی نسبت بالاتری از برچسب‌های تولیدشده برای هر دسته صحیح است (صحت بالاتر).

		PC-HP [19]								
True Class	Browsing	1766	36	46	37	29	30	36	89.2%	10.8%
	Chat	9	417	10	10	3	6	5	90.7%	9.3%
	Email	1	5	278	8	3	6	9	89.7%	10.3%
	File Transfer	2	1	1	134	5	6	1	89.3%	10.7%
	P2P	5	8	4	4	309	4	6	90.9%	9.1%
	Streaming	12	13	12	24	14	809	16	89.9%	10.1%
	VoIP	9	16	12	11	8	15	529	88.2%	11.8%
		97.9%	84.1%	76.6%	58.8%	83.3%	92.4%	87.9%		
		2.1%	15.9%	23.4%	41.2%	16.7%	7.6%	12.1%		
		Browsing	Chat	Email	File Transfer	P2P	Streaming	VoIP		
		Predicted Class								

(الف)

		DISTILLER [17]								
True Class	Browsing	1593	81	68	62	70	59	47	80.5%	19.5%
	Chat	13	382	13	13	16	14	9	83.0%	17.0%
	Email	16	3	251	11	8	11	10	81.0%	19.0%
	File Transfer	5	6	4	118	4	7	6	78.7%	21.3%
	P2P	9	11	10	8	279	8	15	82.1%	17.9%
	Streaming	27	24	24	37	33	732	23	81.3%	18.7%
	VoIP	15	21	15	12	16	19	502	83.7%	16.3%
		94.9%	72.3%	65.2%	45.2%	65.5%	86.1%	82.0%		
		5.1%	27.7%	34.8%	54.8%	34.5%	13.9%	18.0%		
		Browsing	Chat	Email	File Transfer	P2P	Streaming	VoIP		
		Predicted Class								

(ب)

شکل (7). ماتریس درهم‌ریختگی برای طبقه‌بندی ترافیک توسط (الف) روش PC-HP، و (ب) روش DISTILLER

منحنی ویژگی‌های عملیاتی دریافتی^۱ یک نمودار گرافیکی است که برای نشان دادن حساسیت و ویژگی یک الگوریتم طبقه‌بندی برای هر کلاس به‌طور جداگانه استفاده می‌شود. در این منحنی، نرخ‌های مثبت صحیح به‌دست‌آمده از هر کلاس در محور عمودی در برابر نرخ‌های منفی کاذب در محور افقی رسم می‌شود. بدیهی است که طبقه‌بندی کارآمد منجر به نرخ مثبت صحیح بیشتر و نرخ منفی کاذب کمتر می‌شود. بنابراین، مساحت

ماتریس، ستون‌ها با خروجی طبقه‌بندی روش پیشنهادی و سطرها با کلاس واقعی نمونه‌های آزمایشی مطابقت دارند. به‌عنوان مثال، مجموع تمام مقادیر در سطر اول ماتریس نشان می‌دهد که تعداد نمونه‌های دسته‌بندی «Browsing» در طول این آزمایش 1980 مورد بوده که روش پیشنهادی 1897 نمونه از آن‌ها را به‌درستی طبقه‌بندی می‌کند.

		proposed								
True Class	Browsing	1897	6	9	13	11	20	24	95.8%	4.2%
	Chat	4	445	3	2	1	1	4	96.7%	3.3%
	Email		1	289	6	4	4	6	93.2%	6.8%
	File Transfer	2		2	139	4		3	92.7%	7.3%
	P2P	2	4	5	2	320	3	4	94.1%	5.9%
	Streaming	9	12	13	10	4	845	7	93.9%	6.1%
	VoIP	5	3	3	2	8	5	574	95.7%	4.3%
		98.9%	94.5%	89.2%	79.9%	90.9%	96.2%	92.3%		
		1.1%	5.5%	10.8%	20.1%	9.1%	3.8%	7.7%		
		Browsing	Chat	Email	File Transfer	P2P	Streaming	VoIP		
		Predicted Class								

شکل (6). ماتریس درهم‌ریختگی برای طبقه‌بندی ترافیک توسط الگوریتم پیشنهادی

در شکل (7) ماتریس درهم‌ریختگی روش‌های مقایسه شده مشخص شده است. مقایسه ماتریس ارائه‌شده در شکل (6) با ماتریس‌های به‌دست‌آمده از روش‌های مقایسه شده نشان می‌دهد که تعداد خطای طبقه‌بندی روش پیشنهادی برای هر دسته هدف به‌صورت کلی پایین‌تر از سایر روش‌ها است. بررسی مقادیر این ماتریس‌ها بر اساس سطرهای تشکیل‌دهنده آن نشان می‌دهد که نسبت بالاتری از نمونه‌های هر دسته‌ی هدف

¹ Receiver operating characteristic (ROC)

شکل (۸) ب به مقایسه منحنی‌های ویژگی‌های عملیاتی دریافتی برای روش پیشنهادی و سایر روش‌ها می‌پردازد. در این شکل، نقاط آستانه به‌دست‌آمده برای کلاس‌های مختلف باهم تجمیع شده تا عملکرد کلی هر مدل در قالب یک منحنی توصیف گردد. مقایسه نمودارهای ترسیم‌شده در این شکل نشان می‌دهد که روش پیشنهادی قادر به افزایش مساحت زیر نمودار در مقایسه با سایر روش‌هاست. این افزایش از کاهش نرخ مثبت کاذب و افزایش نرخ مثبت صحیح به‌صورت هم‌زمان ناشی می‌شود. لذا می‌توان بر اساس این نمودار، برتری عملکردی مدل پیشنهادی در شناسایی انواع جریان‌های ترافیکی شبکه را نتیجه گرفت.

۵- نتیجه‌گیری

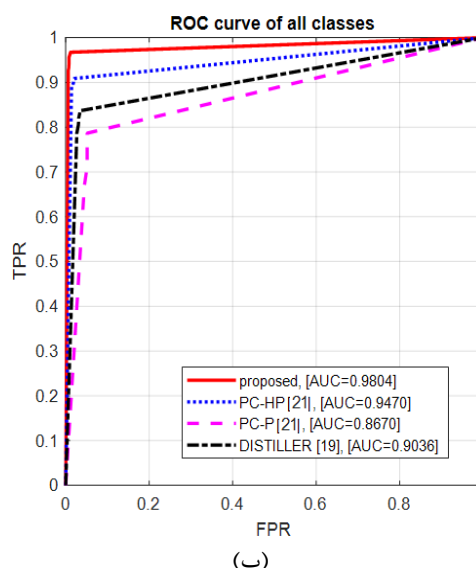
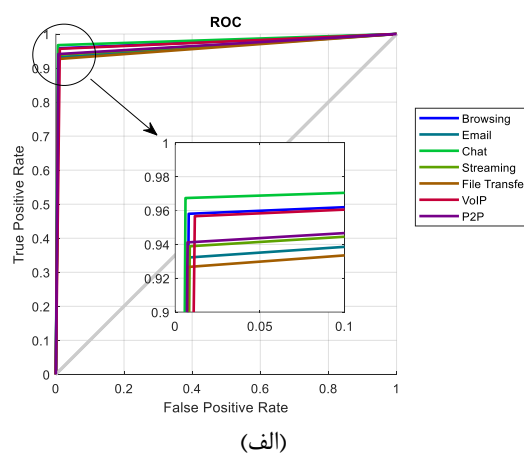
در این تحقیق، یک روش جدید مبتنی بر یادگیری ماشین برای طبقه‌بندی جریان‌های ترافیکی شبکه پیشنهاد گردید. روش پیشنهادی شامل مجموعه‌ای از الگوریتم‌های استخراج ویژگی، خوشه‌بندی و طبقه‌بندی برای شناسایی کلاس صحیح جریان‌های شبکه است. این روش از یک مجموعه کوچک از جریان‌های ترافیکی برچسب‌گذاری شده، در کنار یک مجموعه داده بزرگ که نمونه‌های ترافیک آن بدون برچسب هستند؛ استفاده می‌کند که باعث می‌شود رویکرد پیشنهادی برای کاربردهای واقعی قابل‌استفاده باشد، زیرا در موقعیت‌های واقعی اغلب امکان برچسب‌گذاری تمام نمونه‌های موجود وجود ندارد. رویکرد پیشنهادی، از مجموعه‌ای از ویژگی‌های آماری برای توصیف هر جریان استفاده می‌کند. از آنجایی که ممکن است تعداد ویژگی‌ها زیاد باشد و برخی از آن‌ها به کلاس‌های ترافیکی ربطی نداشته باشند، از الگوریتم تجزیه و تحلیل مؤلفه‌های اصلی برای کاهش ابعاد ویژگی و بهبود سرعت پردازش نمونه‌ها در مراحل بعدی و همچنین کاهش اثر مخرب ویژگی‌های نامرتب استفاده گردید. در ادامه از DBScan برای گروه‌بندی جریان‌های ترافیکی با ویژگی‌های آماری مشابه در خوشه‌های یکسان استفاده شده است. نتایج شبیه‌سازی نشان می‌دهد که روش پیشنهادی قادر به طبقه‌بندی دقیق جریان‌های ترافیکی است و می‌تواند به‌عنوان یک ابزار کارآمد برای سناریوهای دنیای واقعی استفاده شود.

تحقیق حاضر باوجود دستیابی به نتایج مطلوب در زمینه طبقه‌بندی جریان‌های ترافیکی شبکه، با محدودیت‌های زیر مواجه است:

محدودیت‌های داده: کار آیی روش پیشنهادی به کیفیت و کمیت داده‌های موجود بستگی دارد. در مواردی که مجموعه داده‌های برچسب‌گذاری شده محدود یا مغرضانه است، عملکرد مدل ممکن است به خطر بیفتد. اگرچه مدل پیشنهادی با استفاده از راهبرد خوشه‌بندی در جهت رفع این محدودیت تلاش

بیشتر زیر منحنی^۱ نشان کار آیی بیشتر در طبقه‌بندی است. در شکل (۸) الف، منحنی‌های ویژگی‌های عملیاتی دریافتی حاصل‌شده از طریق طبقه‌بندی ترافیک توسط رویکرد پیشنهادی نشان داده شده است. در این شکل منحنی هر کلاس هدف به‌طور جداگانه نشان داده شده است.

طبق این شکل، کلاس "Chat" بالاترین مقدار مساحت زیر منحنی را دارد و کلاس‌های "Browsing" و "Voip" در رتبه بعدی هستند. از طرف دیگر، کمترین مقدار این منحنی متعلق به دسته‌های «Transfer» و «Email» است. به‌طور کلی نتایج این شکل نشان می‌دهد که روش پیشنهادی می‌تواند نمونه‌های هر کلاس را با دقت بالایی طبقه‌بندی کند.



شکل (۸) ب. منحنی‌های ROC به‌دست‌آمده از طریق طبقه‌بندی ترافیک (الف) منحنی به‌دست‌آمده برای رویکرد پیشنهادی به تفکیک هر دسته و (ب) منحنی‌های ROC روش‌های مختلف حاصل از تجمیع خروجی دسته‌های مختلف

[1] M. Bazooband and H. Bahramghiri, "A New Hybrid Approach for Traffic Identification and Classification in Wireless Networks," *Sci. J. Electron. Cyber Def.*, vol. 10, no. 2, 2022. (In Persian). doi: 20.1001.1.23224347.1401.10.2.4.0

[2] S. Z. Majidian, "Reducing the Destructive Effect of Misbehaving Users in Cooperative Spectrum Sensing using Reinforcement Learning," *Sci. J. Electron. Cyber Def.*, vol. 10, no. 4, 2023. (In Persian). doi: 20.1001.1.23224347.1401.10.4.1.1

[3] M. Cotton, L. Eggert, J. Touch, M. Westerlund, and S. Cheshire, "RFC 6335: Internet Assigned Numbers Authority (IANA) Procedures for the Management of the Service Name and Transport Protocol Port Number Registry." RFC Editor, USA, 2011. doi: 10.17487/RFC6335

[4] C. Defense, "Fault Proness Estimation of Software Modules Using Machine learning," *Sci. J. Electron. Cyber Def.*, vol. 11, no. 4, 2024. (In Persian). doi: 20.1001.1.23224347.1402.11.4.4.1

[5] E. Kazemi and A. Taai, "Presenting a New Sentiment Analysis Method Based on Multi-objective Archimedes Optimization Algorithm and Machine Learning," *Sci. J. Electron. Cyber Def.*, vol. 11, no. 4, 2024. (In Persian). doi: 20.1001.1.23224347.1402.11.4.12.9

[6] A. J. R. K. Dadashtabar Ahmadi, M. Kheirkhah, "Detection of advanced Cyber Attacks, Using Behavior Modeling Based on Natural Language Processing," *J. Electron. Cyber Def.*, vol. 3, pp. 141–151, 2018. (In Persian). doi: 20.1001.1.23224347.1397.6.3.12.2

[7] G. Aceto, D. Ciunzio, A. Montieri, and A. Pescapé, "Mobile Encrypted Traffic Classification Using Deep Learning," in 2018 Network Traffic Measurement and Analysis Conference (TMA), 2018, pp. 1–8. doi: 10.23919/TMA.2018.8506558.

[8] J. Yan, "A Survey of Traffic Classification Validation and Ground Truth Collection," in 2018 8th International Conference on Electronics Information and Emergency Communication (ICEIEC), 2018, pp. 255–259. doi: 10.1109/ICEIEC.2018.8473477.

[9] F. Pacheco, E. Exposito, M. Gineste, C. Baudoin, and J. Aguilar, "Towards the Deployment of Machine Learning Solutions in Network Traffic Classification: A Systematic Survey," *IEEE Commun. Surv. Tutorials*, vol. 21, no. 2, pp. 1988–2014, 2019, doi: 10.1109/COMST.2018.2883147.

[10] M. Shen, Y. Liu, L. Zhu, K. Xu, X. Du, and N. Guizani, "Optimizing Feature Selection for Efficient Encrypted Traffic Classification: A Systematic Approach," *IEEE Netw.*, vol. 34, no. 4, pp. 20–27, 2020, doi: 10.1109/MNET.011.1900366.

[11] S. Dong, "Multi class SVM algorithm with active learning for network traffic classification," *Expert Syst. Appl.*, vol. 176, p. 114885, 2021, doi: 10.1016/j.eswa.2021.114885.

[12] Z. Bu, B. Zhou, P. Cheng, K. Zhang, and Z.-H. Ling, "Encrypted Network Traffic Classification Using Deep and Parallel Network-in-Network Models," *IEEE Access*, vol.

کرده است، اما همچنان توانایی مدل در تشخیص دقیق ویژگی‌های متنوع ترافیک شبکه به نمایندگی مجموعه داده مورد استفاده برای آموزش بستگی دارد.

پیچیدگی محاسباتی: پیچیدگی محاسباتی روش پیشنهادی، به‌ویژه مراحل استخراج ویژگی و خوشه‌بندی، برای مجموعه داده‌های بزرگ می‌تواند قابل توجه باشد. اگرچه این پیچیدگی در مقایسه با راهبردهای یادگیری عمیق ناچیز است، اما ممکن است همچنان مقیاس‌پذیری آن را برای کاربردهای بلندپایه یا استقرار بر روی دستگاه‌های دارای منابع پردازشی ناچیز محدود کند.

با توجه به محدودیت‌های ذکر شده و پتانسیل‌های موجود در این حوزه تحقیقاتی، می‌توان پیشنهادهایی را برای ادامه این مسیر تحقیقاتی ارائه داد:

توسعه گام مهندسی ویژگی: توسعه روش‌های مهندسی ویژگی مؤثرتر برای ترافیک شبکه‌های کامپیوتری می‌تواند توانایی مدل را در تشخیص الگوهای مرتبط و کاهش ابعاد فضای ویژگی بهبود بخشد.

بهره‌گیری از هوش مصنوعی قابل توضیح: بررسی روش‌هایی برای قابل توضیح کردن مدل می‌تواند به ایجاد اعتماد به پیش‌بینی‌های آن و تسهیل درک فرآیند تصمیم‌گیری آن کمک کند.

کاربردهای بلندپایه: در تحقیقات آینده می‌توان به چالش‌های محاسباتی موجود به‌منظور استقرار بلندپایه مدل برای طبقه‌بندی ترافیک آنلاین رسیدگی نمود. همچنین بررسی چارچوب‌های محاسباتی توزیع‌شده یا شتاب‌دهنده‌های سخت‌افزاری برای بهبود مقیاس‌پذیری مدل برای رسیدگی به استقرارهای بزرگ‌مقیاس شبکه‌های کامپیوتری می‌تواند موضوعی برای ادامه تحقیقات باشد.

انطباق با الگوهای در حال تکامل ترافیک: توسعه سازوکارهایی برای سازگاری مدل با تغییرات در الگوهای ترافیک در طول زمان، مانند دستگاه‌های نوظهور اینترنت اشیا یا فن‌های جدید حمله می‌تواند تأثیری قابل توجه بر عمومیت کاربرد مدل پیشنهادی داشته باشد.

با رسیدگی به این محدودیت‌ها و دنبال کردن جهت‌گیری‌های پیشنهادشده برای تحقیقات آینده، می‌توان رویکرد پیشنهادی را بیشتر اصلاح و گسترش داد تا چالش‌ها و فرصت‌های در حال تکامل در طبقه‌بندی ترافیک شبکه را برطرف کند.

- Mag., vol. 57, no. 5, pp. 76–81, 2019, doi: 10.1109/MCOM.2019.1800819.
- [21] A. M. Sadeghzadeh, S. Shiravi, and R. Jalili, “Adversarial Network Traffic: Towards Evaluating the Robustness of Deep-Learning-Based Network Traffic Classification,” *IEEE Trans. Netw. Serv. Manag.*, vol. 18, no. 2, pp. 1962–1976, 2021, doi: 10.1109/TNSM.2021.3052888.
- [22] J. Höchst, L. Baumgärtner, M. Hollick, and B. Freisleben, “Unsupervised Traffic Flow Classification Using a Neural Autoencoder,” in *2017 IEEE 42nd Conference on Local Computer Networks (LCN)*, 2017, pp. 523–526. doi: 10.1109/LCN.2017.57.
- [23] F. L. Gewers et al., “Principal Component Analysis: A Natural Approach to Data Exploration,” *ACM Comput. Surv.*, vol. 54, no. 4, May 2021, doi: 10.1145/3447755.
- [24] D. Deng, “DBSCAN Clustering Algorithm Based on Density,” in *2020 7th International Forum on Electrical Engineering and Automation (IFEEA)*, 2020, pp. 949–953. doi: 10.1109/IFEEA51475.2020.00199.
- [25] M. Raja, P. Hasan, M. Mahmudunnobe, M. Saifuddin, and S. N. Hasan. Membership determination in open clusters using the DBSCAN Clustering Algorithm. *Astronomy and Computing*, 47, 100826, 2024. doi: 10.1016/j.ascom.2024.100826.
- [26] J. L. Velazquez-Rodriguez, Y. Villuendas-Rey, O. Camacho-Nieto and C. Yanez-Marquez. A novel and simple mathematical transform improves the performance of lernmatrix in pattern classification. *Mathematics*, 8(5), 732, 2020. doi: /10.3390/math8050732.
- [27] Y. Zhao and X. Shu. Speech emotion analysis using convolutional neural network (CNN) and gamma classifier-based error correcting output codes (ECOC). *Scientific Reports*, 13(1), 20398, 2023. doi: 10.1038/s41598-023-47118-4.
- 8, pp. 132950–132959, 2020, doi: 10.1109/ACCESS.2020.3010637.
- [13] A. A. Afuwape, Y. Xu, J. H. Anajemba, and G. Srivastava, “Performance evaluation of secured network traffic classification using a machine learning approach,” *Comput. Stand. Interfaces*, vol. 78, p. 103545, 2021, doi: 10.1016/j.csi.2021.103545.
- [14] G. D’Angelo and F. Palmieri, “Network traffic classification using deep convolutional recurrent autoencoder neural networks for spatial–temporal features extraction,” *J. Netw. Comput. Appl.*, vol. 173, p. 102890, 2021, doi: 10.1016/j.jnca.2020.102890.
- [15] R. Kumar, M. Swarnkar, G. Singal, and N. Kumar, “IoT Network Traffic Classification Using Machine Learning Algorithms: An Experimental Analysis,” *IEEE Internet Things J.*, vol. 9, no. 2, pp. 989–1008, 2022, doi: 10.1109/JIOT.2021.3121517.
- [16] J. Guan, J. Cai, H. Bai, and I. You, “Deep transfer learning-based network traffic classification for scarce dataset in 5G IoT systems,” *Int. J. Mach. Learn. Cybern.*, vol. 12, no. 11, pp. 3351–3365, 2021, doi: 10.1007/s13042-021-01415-4.
- [17] M. M. Raikar, M. S M, M. M. Mulla, N. S. Shetti, and M. Karanandi, “Data Traffic Classification in Software Defined Networks (SDN) using supervised-learning,” *Procedia Comput. Sci.*, vol. 171, pp. 2750–2759, 2020, doi: 10.1016/j.procs.2020.04.299.
- [18] G. Aceto, D. Ciunzo, A. Montieri, and A. Pescapé, “Toward effective mobile encrypted traffic classification through deep learning,” *Neurocomputing*, vol. 409, pp. 306–315, 2020, doi: 10.1016/j.neucom.2020.05.036.
- [19] G. Aceto, D. Ciunzo, A. Montieri, and A. Pescapé, “DISTILLER: Encrypted traffic classification via multimodal multitask deep learning,” *J. Netw. Comput. Appl.*, vol. 183–184, p. 102985, 2021, doi: 10.1016/j.jnca.2021.102985.
- [20] S. Rezaei and X. Liu, “Deep Learning for Encrypted Traffic Classification: An Overview,” *IEEE Commun.*