Scientific Journal of Electronic & Cyber Defense
Vol. 12, No. 4, Winter 2025, Serial No. 48 , pp.95-105
ISSN: 2322-4347, E-ISSN: 2980-8979

Violent behavior detection in surveillance cameras using convolutional and memory neural networks

A. Dolat Khah, M. Asadpour, R. Hashempour, B. Dorostkar*

*Assistant Professor, Amin University, Tehran, Iran

(Received: 2024/08/21, Revised: 2024/12/08, Accepted: 2025/01/02, Published: 2025/02/01)
DOR: https://dor.isc.ac/dor/20.1001.1.23224347.1403.12.4.9.3

Abstract

The existence of security is mandatory in any society and it is the basis for the progress and development of a country as easily and quickly as possible, so all countries try to establish stable security by controlling the level of violence and strife in the society. On the other hand, due to the limitation of manpower, it is not possible to carry out the entire process of providing security through the traditional and common methods of the past, and in this regard, new and up-to-date equipment and technologies must be used. and advanced countries of the world, the use of closed-circuit and surveillance cameras in public places is in this research, an expert system based on two sets of neural networks ResNet101 and memory LSTM to reduce the amount of computation while maintaining proper accuracy, ResNet101 network is presented With a total of 347 layers and through the transfer learning method, it extracts the spatiotemporal features of consecutive video frames, and then the LSTM network with a total of 9 layers is responsible for detecting violent behavior in the video. These two sets have been optimized in terms of the type of layer arrangement, the way of connection, and the number of cells in each layer so that they can have the best performance in all video conditions, including low quality, presence of noise short video, etc. As a result of this intelligent system, they can detect violent behavior in closed-circuit cameras with an accuracy of 86.28% in real-time and instantaneously in low-quality video images of 224x3x224 pixels, and cases of violence, report it to the relevant people. inform In the end, it should be mentioned that the designed system, by reducing the amount of computing while maintaining accuracy, has been able to perform effective and appropriate online monitoring of low-quality surveillance cameras by using only 22 frames per 5 seconds of video.

Keywords: violence detection, machine vision, artificial neural networks, surveillance cameras

Cite this article: A. Dolat Khah, M. Asadpour , R. Hashempour, B. Dorostkar “A method for software fault prediction with combination of chaos theory and herd horse optimization algorithm,” Electronic and Cyber Defense, vol.12 , no.4 , pp.95-105 , .
DOR: https://dor.isc.ac/dor/20.1001.1.23224347.1403.12.4.9.3
 Publisher: Imam Hossein University

*Corresponding Author Email: behnamdorostkar@gmail.com

"پدافند الکترونیکی و سایبری"
سال دوازدهم، شماره ۴، زمستان ۱۴۰۳، ص ۱۰۵-۹۵
شاپا الکترونیکی: 2980-8979 شاپا چاپی: 2322-4347

علمی – پژوهشی

تشخیص رفتارهای خشونت‌آمیز در دوربین‌های نظارتی به کمک شبکه‌های عصبی پیچشی و حافظه‌دار

احمد دولت‌خواه[۱]، میلاد اسد پور[۲]، راهب هاشم‌پور[۳]، بهنام درستکار[۴*]

۱- مربی، ۲- دانشجوی دکتری، ۳- کارشناسی ارشد، ۴- استادیار، جامع علوم انتظامی امین، تهران، ایران

(دریافت: ۱۴۰۳/۰۵/۳۰، بازنگری: ۱۴۰۳/۰۹/۱۸، پذیرش: ۱۴۰۳/۱۰/۱۳، انتشار: ۱۴۰۳/۱۱/۱۳)

DOR: https://dor.isc.ac/dor/20.1001.1.23224347.1403.12.4.9.3

چکیده

وجود امنیت در هر جامعه‌ای الزامی است و زمینه‌ساز پیشرفت و توسعه هر چه راحت‌تر و سریع‌تر یک کشور است؛ لذا تمامی کشورها سعی در برقراری امنیتی پایدار از طریق کنترل میزان خشونت و نزاع در سطح جامعه دارند. از طرفی به دلیل محدودیت نیرو انسانی نمی‌توان تمامی فرایند تأمین امنیت را از طریق روش‌های سنتی و متداول گذشته انجام داد و باید در این راستا از تجهیزات و فنّاوری‌های جدید و بروز دنیا استفاده کرد یکی از این فناوری که اخیراً بسیار موردتوجه جامعه بین‌الملل و کشورهای پیشرفته دنیا قرارگرفته است استفاده از دوربین‌های مداربسته و نظارتی در اماکن عمومی است در این پژوهش سیستمی خبره بر اساس دو مجموعه شبکه عصبی ResNet101 و حافظه‌دار LSTM باهدف کاهش حجم محاسباتی در عین حفظ دقت مناسب، ارائه‌شده است که شبکه ResNet101 با مجموع ۳۴۷ لایه و از طریق روش یادگیری انتقال ویژگی‌های فضا-زمانی فریم‌های متوالی ویدئو را استخراج نموده و سپس شبکه LSTM با مجموع ۹ لایه، وظیفه تشخیص رفتار خشونت‌آمیز در ویدئو را بر عهده دارد. این دو مجموعه ازنظر نوع چینش لایه، نحوه اتصال و تعداد سلول در هر لایه به‌گونه‌ای بهینه‌شده‌اند که بتوانند در تمامی شرایط ویدئویی اعم از کیفیت پایین، وجود نویز و کوتاهی ویدئو و... بهترین عملکرد را داشته باشند. درنتیجه این سامانه هوشمند می‌توانند با دقت ۸۶/۲۸٪ به‌صورت بلادرنگ و لحظه‌ای در تصاویر ویدئویی باکیفیت پایین ۳*۲۲۴*۲۲۴ پیکسلی به تشخیص رفتارهای خشونت‌آمیز در دوربین‌های مداربسته بپردازند و در صورت وقوع خشونت مراتب آن را به افراد ذی‌ربط اطلاع دهند. در انتها باید اشاره داشت که سیستم طراحی‌شده با کاهش حجم محاسباتی در کنار حفظ میزان دقت توانسته است فقط با استفاده از ۲۲ فریم در هر ۵ ثانیه از ویدئو، پایش کارا و مناسبی را به‌صورت برخط در دوربین‌های نظارتی باکیفیتی پایین انجام دهد.

کلیدواژه‌ها: تشخیص خشونت، بینایی ماشین، شبکه‌های عصبی مصنوعی، دوربین‌های نظارتی

۱. مقدمه

از ابتدای تاریخ تمدن بشری تاکنون بحث امنیت جوامع بشری همواره مورد تأکید تمامی حکومت‌ها و جوامع بوده است. انسان همیشه تمایل دارد که در محیطی آرام و به‌دوراز خشونت و هرج مرج بتواند به فعالیت‌های شخصی و اجتماعی خود برسد تا از این طریق علاوه بر رشد و سعادت شخصی بتواند به تعالی و پیشرفت جامعه کمک نماید؛ لذا حکومت‌ها در هر زمان و مکانی سعی بر آن دارند که میزان خشونت در جامعه را کنترل نمایند تا از این طریق فضایی باثبات و با امنیت پایدار ایجاد نمایند. یکی از اصول اساسی امینت تشخیص و جلوگیری از خشونت است. درصورتی‌که خشونت در یک جامعه‌ای به‌درستی و در زمان حداقلی کشف و شناسایی گردد و در سریع‌ترین زمان ممکن با متخلفین برخورد قانونی گردد، می‌توان ایجاد امنیتی پایدار و گسترده را در جامعه تا حدودی تضمین نمود؛ لذا یکی از مهم‌ترین زمینه‌های موردبحث برای حکومت‌ها شناسایی به‌موقع و سریع خشونت رخداده در جامعه است از گذشته‌های بسیار دور تا حدوداً یک

استناد: دولت‌خواه، احمد، اسدپور، میلاد، هاشم‌پور، راهب، درستکار، بهنام " تشخیص رفتارهای خشونت‌آمیز در دوربین‌های نظارتی به کمک شبکه‌های عصبی پیچشی و حافظه‌دار"، پدافند الکترونیکی و سایبری، ۱۲(۴)، ۱۰۵-۹۵، ۱۴۰۳.
https://dor.isc.ac/dor/20.1001.1.23224347.1403.12.4.9.3

ناشر: دانشگاه جام امام حسین(ع).

* رایانامه نویسنده مسئول: behnamdorostkar@gmail.com

قرن گذشته تشخیص و شناسایی خشونت بر عهده سازمان‌های امنیتی کشور همچون پلیس قرار داشته است و پلیس در هر کشوری متعهد به جلوگیری از خشونت در هر جامعه‌ای بود با افزایش یک‌باره و انفجاری جمعیت جهان که در اثر پایان جنگ‌های جهانی و قحطی‌های پی‌درپی اتفاق افتاد بود، کنترل جمعیت وسیعی از افراد به‌صورت سنتی و توسط پلیس بسیار سخت و غیرممکن شد درنتیجه استفاده گسترده از تجهیزات جانبی اعم از دوربین‌های نظارتی و مداربسته در اماکن عمومی گسترش چشمگیری یافت از طرفی استفاده از دوربین‌های مداربسته نیز چالش‌هایی با خود به همراه داشت؛ زیرا این دوربین‌ها به‌صورت ۲۴ ساعته و شبانه‌روزی باید پایش و کنترل می‌شدند و باتوجه‌به تعداد بالایی دوربین‌ها امکان نظارت بر تمامی آن‌ها بسیار سخت و غیرممکن شده بود. با پیشرفت علوم هوش مصنوعی و شبکه‌های عصبی مصنوعی استفاده از روش‌های کامپیوتری و هوشمند در علوم مختلف رواج پیدا کرد یکی از این علوم پردازش تصویر و ویدئو است روش‌های مختلفی نیز در این حوزه بر اساس علوم و فنون پردازش تصویر و هوش مصنوعی ارائه‌شده است که عمده این تجهیزات و روش‌ها زمان‌بر بوده و از دقت مناسبی نیز برخوردار نیستند و کیفیت پایین تصاویر نیز تأثیر بسزایی بر روی کاهش عملکرد آنان می‌گذارد؛ لذا استفاده عملیاتی از این روش‌ها در جامعه عموماً امکان‌پذیر نیست. در این پژوهش سعی شده است با استفاده روش‌های هوش مصنوعی و بینایی ماشین مشکل ذکرشده در کنترل و نظارت شبانه‌روزی بر دوربین‌های مداربسته رفع گردد. در این پژوهش از یک شبکه از پیش آموزش‌دیده شده ResNet 101[1] به همراه یک شبکه عصبی حافظه‌دار LSTM[2] برای تشخیص و شناسایی بلادرنگ و به‌موقع رفتارهای خشونت‌آمیز در تصاویر دوربین‌های مداربسته استفاده‌شده است. این شبکه می‌تواند بلادرنگ و با دقت ٪۸۳ به بررسی تصاویر دوربین‌های نظارتی بپردازد و در صورت بروز هرگونه خشونت در تصاویر مراتب را به مأمورین ذی‌ربط ارائه دهد.

باتوجه‌به تعداد زیاد دوربین نظارتی در سطح جامعه و همچنین تعداد بالای رخداد خشونت در طی ۲۴ ساعت شبانه‌روز و از طرفی تعداد محدود کارکنان امنیتی، پایش تصاویر به‌صورت سنتی و توسط اشخاص حقیقی عملاً غیرممکن است لذا باید سیستم‌های هوشمندی طراحی گردد که بتواند وظیفه تشخیص خشونت و رفتارهای غیرعادی را به‌صورت خودکار و برخط و سریع انجام دهد و در صورت وجود هرگونه خشونت در دوربین‌ها، ویدیو خشونت رخ‌داده شده، ذخیره گردیده و مراتب به مأمورین ذی‌ربط اطلاع داده شود.

در این مقاله تلاش شده است تا باتوجه‌به چالش بیان‌شده با ایجاد نوآوری در ساختار و.... به بهبود نتایج دست‌یافته که در ادامه به‌صورت کامل تشریح خواهد شد.

ساختار مقاله بدین شرح است که در بخش دوم پیشینه موردبررسی قرار می‌گیرد. بخش سوم روش پیشنهادی و معماری شبکه مربوطه تحلیل خواهد شد. نتایج پیاده‌سازی در نرم‌افزار در بخش پنجم ارائه می‌گردد.

۲. روش تحقیق

با گسترش رفتارهای خشونت‌آمیز در جوامع امروزی تشخیص به‌موقع و بلادرنگ این‌گونه رفتارها موردتوجه ویژه سازمان‌های امنیتی قرار گرفت از طرفی این کار با سختی‌های زیادی نیز همراه بود که ازجمله آن می‌توان به کمبود نیروی انسانی و عدم توانایی افراد برای بازرسی ۲۴ ساعته اشاره کرد؛ لذا با گسترش دوربین‌های مداربسته و فنّاوری پایش تصاویر و پیشرفت روزافزون علم هوش مصنوعی و بینایی ماشین پژوهشگران درصدد استفاده از این علوم برای تشخیص خودکار و برخط رفتارهای خشونت‌آمیز در دوربین‌های نظارتی برآمدند و تحقیقات بسیار در این زمینه صورت گرفت و روش‌های گوناگون و متنوعی نیز در این راه استفاده گردید که در ادامه به تعدادی از این پژوهش‌ها پرداخته‌شده است.

گارسیا و همکاران[۱] طی پژوهش خود از ویژگی‌های مختلف زمانی و مکانی مانند شتاب حرکت و تشخیص توده در فریم‌های متوالی تصویر استفاده کردند. علی‌رغم نتایج دلگرم‌کننده‌ای که در آن نرخ‌های دقت بالایی به دست آمد، هزینه محاسباتی استخراج چنین ویژگی‌هایی برای کاربردهای عملی گران است و از آن برای تشخیص بلادرنگی نمی‌توان استفاده کرد. این روش برای مجموعه داده‌های مختلف دقت‌های متفاوتی را دارد که دقت آن ٪۸۰ بر روی مجموع مجموعه داده‌های خشونتی است.

فیلم‌ها به یکی از منابع اصلی سرگرمی در عصر حاضر تبدیل‌شده‌اند که بر اساس ایده‌های متنوعی شکل‌گرفته‌اند . فیلم‌های اکشن در چند سال اخیر بیشترین توجه را به خود جلب کرده‌اند که حاوی صحنه‌های خشن هستند، زیرا یکی از ویژگی‌های نامطلوب برای برخی افراد است که برای ایجاد جذابیت و فانتزی استفاده می‌شود .بااین‌حال، این صحنه‌های خشونت‌آمیز تأثیر منفی روی بچه‌ها گذاشته است و حتی برای افراد بالغ هم راحت نیست .بهترین راه برای جلوگیری از تماشای صحنه‌های خشن در فیلم‌ها توسط افراد زیر سن، حذف این صحنه‌ها است .در پژوهش سامی اولاه خان [۲] یک طرح

[1] Resiual Network.
[2] Long short term memory.

تشخیص خشونت برای فیلم‌ها پیشنهاد شد که شامل سه مرحله است. ابتدا کل فیلم به پلان‌ها تقسیم می‌شود و سپس یک فریم نماینده از هر پلان بر اساس سطح برجسته انتخاب می‌شود. سپس، این فریم‌های انتخاب‌شده از یک مدل یادگیری عمیق سبک‌وزن منتقل می‌شوند که با استفاده از رویکرد یادگیری انتقالی برای طبقه‌بندی نماهای خشونت‌آمیز و بدون خشونت در یک فیلم به‌خوبی تنظیم‌شده است. درنهایت، تمام صحنه‌های بدون خشونت در یک سکانس ادغام می‌شوند تا فیلمی عاری از خشونت تولید شود که می‌تواند توسط کودکان و همچنین افراد پارانویید خشونت دیده شود. مدل پیشنهادی بر روی سه مجموعه داده معیار خشونت ارزیابی می‌شود و به‌طور تجربی ثابت می‌شود که طرح پیشنهادی تشخیص سریع و دقیق صحنه‌های خشونت‌آمیز در فیلم‌ها را در مقایسه با روش‌های پیشرفته ارائه می‌دهد. در مقاله دیگری از پنگ وو و همکاران [۳]ابتدا یک مجموعه داده در مقیاس بزرگ و چند صحنه به نام XD-Violenceبا مدت‌زمان ۲۱۷ ساعت، حاوی ۴۷۵۴ ویدئوی پاک نشده با سیگنال‌های صوتی و برچسب‌های ضعیف منتشر کردند. سپس یک شبکه عصبی حاوی سه‌شاخه موازی را پیشنهاد شد تا روابط مختلف بین قطعه‌های ویدئویی را ثبت کند و ویژگی‌ها را یکپارچه کند که در آن شاخه کل‌نگر وابستگی‌های دوربرد را با استفاده از شباهت قبلی به‌صورت پویا ثبت می‌کند. روش آنان علاوه بر داده‌های ویدئویی از داده‌های صوتی داخل تصاویر نیز استفاده می‌کند؛ لذا نتایج حاصل از این پژوهش در مقایسه با دیگر پژوهش‌ها از دقت بالاتری برخوردار است و دقت ۹۵٪ را ارائه می‌کند. در پژوهشی دیگر از محمودی و همکاران [۴]یک توصیفگر ویژگی جدید و بدیع به نام هیستوگرام اندازه و جهت جریان نوری[1]را معرفی کردند که ابتدا، روش پیشنهادی فریم‌های ورودی را به فرمت مقیاس خاکستری تبدیل می‌کند. سپس، جریان نوری بین دو فریم پیاپی را محاسبه می‌کند. سپس، مقدار جریان نوری و جهت هر پیکسل در هر فریم به‌طور جداگانه با قاب قبلی خود مقایسه می‌شود تا تغییرات معنی‌داری در اندازه و جهت به دست آید. پس‌ازآن، مقادیر آستانه متفاوتی برای تغییرات بزرگی و جهت‌گیری برای به دست آوردن شش شاخص باینری اعمال می‌شود .درنهایت، این شاخص‌های باینری برای به دست آوردن توصیفگر HOMO که برای آموزش یک طبقه‌بندی کننده SVM استفاده می‌شود، تجزیه‌وتحلیل می‌شوند. این سیستم با استفاده از متلب پیاده‌سازی شده است. برای ارزیابی روش پیشنهادی، از دو مجموعه داده معیار استفاده‌شده است. مقایسهHOMO و سایر توصیفگرها در مجموعه داده‌های معیار عملکرد رضایت‌بخشی را نشان می‌دهد.

در پژوهشی از فبین و همکاران [۵]از مجموعه‌ای غنی از ابزارها را برای تجزیه‌وتحلیل اطلاعات حرکتی غیرمحلی و بلندمدت ارائه می‌کند، استفاده کردند. بر اساس این روش یک فن لاگرانژی تخصصی برای تشخیص خودکار صحنه‌های خشونت‌آمیز در فیلم‌های ویدئویی ارائه شد. در این پژوهش از فیلترهای حرکتی و زمانی نیز برای تشخیص دقیق‌تر خشونت استفاده‌شده است و برای اطمینان از مقیاس‌های ویژگی‌های مکانی و زمانی مناسب، یک روش توسعه‌یافته بر اساس تئوری bag-of-words به‌عنوان طرح طبقه‌بندی بر اساس هر ویدئو اعمال شد که این مورد به افزایش دقت و عملکرد روش کمک شایانی نمود این روش باتوجه‌به پارامتر زمانی t و فواصلی که برای مقیاس در نظر گرفته می‌شود. از دقت‌های متفاوتی بهره می‌برد که می‌توان گفت به‌طور میانگین دقت این روش ٪۸۳/۳۴ است. البته باید توجه داشت این روش بلادرنگ و بر خط نبوده است و حجم محاسباتی بالایی را شامل می‌شود و از ویدئوهایی باکیفیت بالا استفاده می‌نماید.

در سال‌های اخیر، دوربین‌های نظارتی به‌طور گسترده در مکان‌های عمومی مستقرشده‌اند و میزان جرم عمومی به دلیل این دستگاه‌های همه‌جا به میزان قابل‌توجهی کاهش‌یافته است. معمولاً این دوربین‌ها پس‌از انجام جنایات نشانه‌ها و شواهدی را ارائه می‌دهند، درحالی‌که به‌ندرت برای جلوگیری یا توقف به‌موقع فعالیت‌های مجرمانه استفاده می‌شوند. نظارت دستی مقدار زیادی از داده‌های ویدئویی از دوربین‌های نظارتی زمان زیادی را می‌طلبد؛ بنابراین، تشخیص خودکار رفتارهای خشونت‌آمیز از سیگنال‌های ویدئویی ضروری است. پژوهش پراتاما و همکاران [[چندین مجموعه داده ویدئویی موجود را برای تشخیص خشونت خلاصه می‌کند و پایگاه داده RWF-2000 را با ۲۰۰۰ ویدئوی ضبط‌شده توسط دوربین‌های نظارتی درصحنه‌های دنیای واقعی پیشنهاد می‌کند. همچنین، یک روش جدید و نوین ارائه‌شده است که از مزایای شبکه‌های 3D-CNN و جریان نوری به نام Flow Gated Network استفاده می‌کند. روش پیشنهادی دقت ٪۸۷/۲۵ را در مجموعه آزمایشی پایگاه داده پیشنهادی به دست می‌آورد. پایگاه داده و کدهای منبع در حال حاضر برای دسترسی، باز هستند این پژوهش بلادرنگ نبوده و از حجم محاسباتی بالایی برخوردار است.

در پژوهشی از عبدالرحمان ترائوره و همکاران [۷]یک معماری پیشنهاد جدید عصبی پیشنهادشده است که یک واحد بازگشتی دروازه‌ای دوطرفه[2] و یک شبکه عصبی پیچشی دوبعدی را برای تشخیص خشونت در دنباله‌های ویدئویی ترکیب می‌کند. یک CNNبرای استخراج ویژگی‌های فضایی از هر فریم استفاده

[1] Flow rate.

[2] LSTM Layer.

می‌شود، درحالی‌که BiGRU ویژگی‌های حرکت زمانی و محلی را با استفاده از ویژگی‌های استخراج‌شده CNN از چندین فریم استخراج می‌کند. شبکه یادگیری عمیق پیشنهادی در سه مجموعه داده عمومی با پیچیدگی‌های صحنه متفاوت آزمایش می‌شود. شبکه پیشنهادی تا ۹۸ درصد دقت دارد. نتایج به‌دست‌آمده امیدوارکننده هستند و عملکرد رویکرد پیشنهادی را نشان می‌دهند؛ ولی باید توجه داشت که این روش بر روی مجموعه داده نزاع در ورزش هاکی صورت گرفته است که دارای کیفیتی بالا با صحنه‌هایی بدیهی و واضح از نزاع است؛ لذا مقایسه نتایج حاصل از نتیجه در مقایسه با پژوهش فعلی زیاد درست نیست.

به‌طورکلی، تشخیص خشونت در یک محل شلوغ، از طریق رویکردهای محاسباتی، به دلیل حرکات سریع، ویژگی‌های همپوشانی، مناظر مسدود و پس‌زمینه‌های پراکنده چالش‌برانگیز است .خوشبختانه، فن‌های یادگیری عمیق می‌توانند ناهنجاری‌ها را تا حدودی شناسایی کنند .علاوه بر این، محبوبیت آن‌ها، به‌عنوان الگوی تشخیص خشونت، با سرعت فوق‌العاده‌ای در حال افزایش است .هدف از چنین رویکردهایی توسعه روشی است که خشونت را تشخیص دهد و زنگ خطر را برانگیزد تا بتوان کمک فوری ارائه کرد. چادهاری و همکاران[۹]باتوجه‌به افزایش روزافزون جرم و جنایت روشی را برای تشخیص رفتار در دوربین ارائه کردند این روش فعالیت رخداده در دوربین‌های نظارتی را به‌صورت خودکار شناسایی و طبقه‌بندی می‌کند سپس تعیین می‌نماید که این فعالیت جزو رفتارهای خشونت‌آمیز است یا خیر. چارچوب پیشنهادی چادهاری و همکاران شامل سه مرحله اصلی است: تشخیص شی متحرک، ردیابی شیء و درک رفتار برای تشخیص فعالیت. با استفاده از فرآیند استخراج ویژگی، ویژگی‌های کلیدی (سرعت، جهت، مرکز و ابعاد شیء موردنظر) شناسایی می‌شوند. این ویژگی‌ها به ردیابی شیء در فریم‌های ویدئویی کمک می‌کند در انتها نیز یک طبقه‌بندی کننده فعالیت تعیین می‌نماید که آیا فعالیت صورت گرفته در تصویر خشونت‌آمیز است یا خیر. در پژوهش چادهاری و همکاران اشاره‌شده است که این روش با موفقیتی ۹۰٪ می‌تواند فعالیت‌های خشونت‌آمیز در فیلم‌های سینمایی تشخیص دهد. دینگ و همکاران[۱۰]برخلاف تحقیقات پیشین از یک شبکه ConvNet سه‌بعدی از قبل آموزش‌دیده شده به‌منظور شناسایی و تشخیص رفتار خشونت‌آمیز در ویدئو استفاده کردند. این شبکه به‌صورت خودکار ویژگی‌های مربوط به تصاویر خشونت‌آمیز را استخراج نموده و سپس با اصلاح وزن‌ها و بایاس‌های داخلی شبکه میزان خطای شبکه را به حداقل می‌رساند؛ لذا پس‌ازآن شبکه می‌توان ورودی ویدئویی جدید را شناسایی و بر اساس دانسته‌های خود که از مجموعه داده‌های آموزشی به دست آورد است دریابد که آیا رفتار خشونت‌آمیزی رخداده است یا خیر. در شکل 7 می‌توان نمای کلی شبکه ConvNet سه‌بعدی که دینگ و همکاران از آن استفاده نمودند به همراه تعداد سلول‌های موجود در هر لایه را مشاهده نمود. نتایج حاصل از این تحقیق نشان می‌دهد که روش ارائه‌شده توسط دینگ در بهترین حالت و بدون درنظرگرفتن حجم محاسباتی می‌تواند با دقت ۸۹٪ رفتار خشونت‌آمیز را در ویدئو تشخیص دهد البته باید توجه داشت که دقت شبکه باتوجه‌به ضریب یادگیری مقادیر متفاوتی خواهد داشت و این دقت برای دوربین‌های نظارتی باکیفیت پایین ۷۹٪ است. در پژوهش سانتوس و همکاران [۱۱]یک مطالعه تطبیقی از معماری‌های پیشرفته یادگیری عمیق را ارائه می‌کند که برای تشخیص خشونت در داخل خودرو فقط بر اساس سیگنال صوتی اعمال می‌شوند. روش پیشنهادی برای نمایش سیگنال صوتی Mel-spectrogramبود، پس از بررسی عمیق ادبیات گذشته در این حوزه چهار معماری مختلف یادگیری عمیق برای حل مشکل طبقه‌بندی اعمال شد. نتایج نشان داده است که مدلResNet-18 بهترین نتایج دقت را در مجموعه آزمایشی ارائه می‌دهد و دقتی برابر ۸۲٪ درصد ارائه نموده است.

رفتارهای خشونت‌آمیز، نزاع و درگیری‌های خیابانی همواره تأثیرات بسیار مخربی در اقتصاد و امنیت جامعه داشته‌اند و افزایش بیش‌ازحد آن می‌تواند یک جامعه را به ورطه نابودی برساند؛ لذا همواره و در تمامی جوامع پیشرفته امروزی تشخیص رفتار خشونت‌آمیز یک اصل اساسی و مهم در امنیت پایدار جامعه محسوب می‌شود و دولت‌ها همواره در تلاش‌اند که هر چه سریع‌تر و دقیق‌تر بتواند رفتارهای خشونت‌آمیز در جامعه را تشخیص داده و با عامل آن برابر قانون برخورد نمایند در اینجا لازم به توضیح است که یکی از عوامل اصلی رضایت یک جامعه از دولت حاکم بر آن نیز مسئله امنیت پایدار، آسایش و راحتی است که خود در گروه تشخیص به‌موقع و برخورد قاطع با رفتارهای خشونت‌آمیز و نزاع و درگیری است. یکی از مهم‌ترین این روش‌ها برای افزایش امنیت شناسایی و تشخیص دقیق و به‌موقع خشونت با استفاده دوربین‌های نظارتی است. در این پژوهش با استفاده از شبکه عصبی پیچشی ResNet و همچنین شبکه عصبی حافظه‌دار LSTM سیستمی هوشمند طراحی و پیاده‌سازی شده است که می‌تواند علی‌رغم وجود ویدئوهایی باکیفیت پایین و نویز دار به‌صورت بلادرنگ و بر خط به تشخیص خشونت در دوربین‌های مداربسته بپردازد. این سیستم به‌گونه‌ای طراحی‌شده است که بتواند فقط با استفاده از ۲۲ فریم در ۵ ثانیه وظیفه تشخیص خشونت در دوربین‌ها انجام دهد؛ لذا این سامانه توانسته است در کنار کاهش چشمگیر حجم محاسبات با دقت مناسبی به تشخیص خشونت در دوربین‌های مداربسته بپردازد.

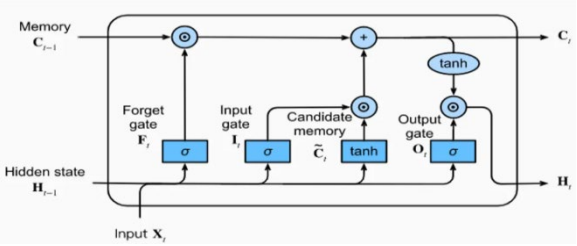
۲-۲. شبکه‌های عصبی ResNet

شبکه رزنت یا به‌صورت کامل، Residual Network یکی از شبکه‌های عمیق معروف در سطح جهان است. این شبکه را شِیوکینگ رن، کِیمینگ هی، ژان سان و زایانگیا ژوئنگ در سال ۲۰۱۵ معرفی کردند[۱۲]. مدل رزنت تاکنون یکی از محبوب‌ترین و موفق‌ترین مدل‌های یادگیری عمیق بوده است. دلیل موفقیت شبکه رزنت این است که امکان آموزش شبکه‌های عصبی بسیار عمیق با بیش از ۱۵۰ لایه را بدون ایجاد مشکل خاصی در فرآیند آموزشی دارد. قبل از رزنت شبکه‌های عصبی بسیار عمیق، به دلیل مشکل محوشدگی گرادیان[۱] ، دچار مشکل می‌شدند.

شبکه‌های یادگیری عمیق معمولی، مانند AlexNet ، ZFNet و VGGNe، اغلب لایه‌های پیچشی و سپس لایه‌های کاملاً متصل[۲] برای طبقه‌بندی دارند که بدون هیچ‌گونه اتصال میان‌بر به یکدیگر متصل شده‌اند و همین اتصال مستقیم می‌تواند باعث اختلال در فرآیند آموزش شبکه شود. اتصالات میانبر[۳] یا اتصالات اضافی[۴] راه‌حلی بود که شبکه رزنت برای حل مشکل شبکه‌های عمیق ارائه کرد. این شبکه با دارا بودن 84.8 میلیون پارامتریکی از پیچیده‌ترین و دقیق شبکه‌های عصبی دنیا به‌حساب می‌آید.[۱۳]

۲-۳. شبکه‌های عصبی LSTM

شبکه بازگشتی یا RNN[۵] نوعی شبکه عصبی است که حافظه داخلی دارد؛ به‌عبارت‌دیگر، این شبکه یک شبکه عصبی معمولی است که در ساختارش حلقه‌ای دارد که از طریق آن در هر گام خروجی، گام قبلی به همراه ورودی جدید، به شبکه وارد می‌شود. این حلقه به شبکه کمک می‌کند تا اطلاعات قبلی را در کنار اطلاعات جدید داشته باشد و بتواند بر اساس این اطلاعات خروجی مدنظر را تولید نماید. این ویژگی بسیار مناسب شبکه‌های بازگشتی این امکان را می‌دهد که بتوان از آن‌ها در داده‌های ترتیبی[۶] ، مانند متن، صدا و تصویر استفاده نمود. شبکه LSTMنوع خاصی از شبکه RNN است که مشکل حافظه بلندمدت شبکه RNN را حل می‌کند. شبکهLSTM سازوکارهایی داخلی به اسم گیت[۷] دارد. این گیت‌ها جریان اطلاعات را کنترل می‌کنند؛ همین‌طور مشخص می‌کنند چه داده‌هایی در توالی مهم هستند و باید همچنان حفظ بشوند و چه داده‌هایی باید حذف بشوند؛ به این شکل، شبکه اطلاعات مهم را در طول زنجیره توالی عبور می‌دهد تا خروجی ایجاد گردد[۱۴]شکل ۱ نمایی از یک سلول LSTM را نشان می‌دهد.



شکل(۱): نمایی از لایه LSTM

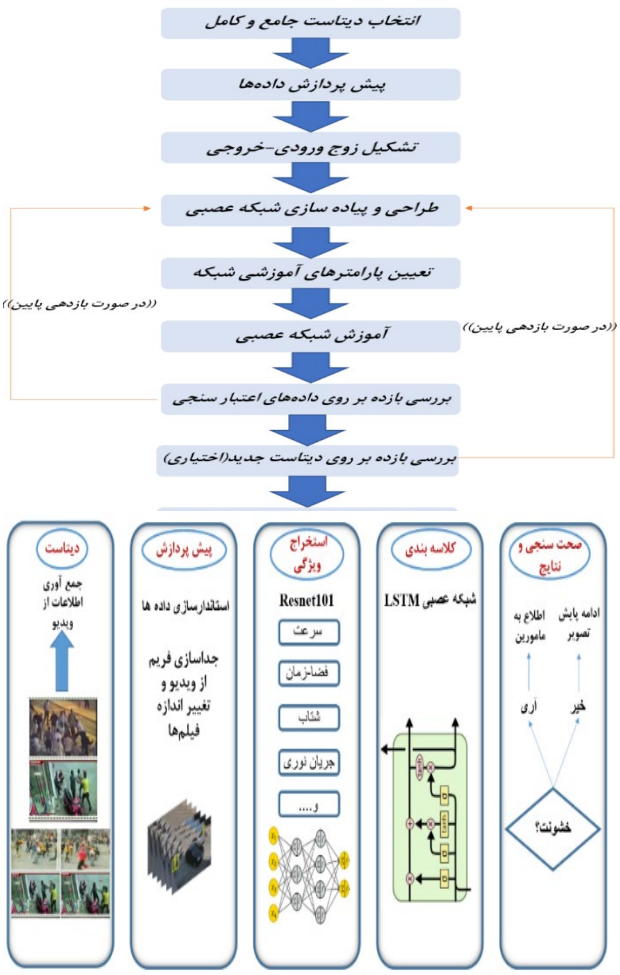
۲-۴. روش انجام‌شده در پژوهش

برای انجام پژوهش در حوزه هوش مصنوعی و شبکه‌های عصبی در ابتدا نیازمند یک مجموعه کامل و جامع از دادگان است که از آن با عنوان مجموعه داده[۸] یاد می‌گردد. این مجموعه دادگان توسط دانشمندان علوم داده و به‌صورت دقیق و جامع و با رعایت اصول و استانداردها در اختیار پژوهشگران دنیا قرار می‌گیرد. پس از انتخاب مجموعه داده و فراهم آوردن آن باید مجموعه دادگان موجود بر اساس روش‌های استانداردسازی[۹]، به داده‌هایی مناسب با توزیع گوسی تبدیل گردد که ازجمله این روش‌ها می‌توان به z-score normalization و min-max normalization اشاره کرد که این دو روش ازجمله متداول‌ترین روش‌ها برای انجام این کار است. به این مرحله که داده‌ها استانداردسازی و برای ورود به شبکه اصلی سازگار می‌شوند پیش‌پردازش اطلاعات گفته می‌شود. پس از پیش‌پردازش اطلاعات مجموعه داده بر اساس برچسب‌ها باید به زوج ورودی-خروجی تقسیم گردد. بدین‌صورت که شبکه برای هر داده ورودی بتواند خروجی مطلوب آن را نیز در نظر بگیرد. این کار برای یادگیری شبکه لازم و ضروری است پس‌ازاین مهم باید شبکه عصبی متناسب با وظیفه‌ی محول شده به هوش مصنوعی طراحی و پیاده‌سازی گردد این شبکه از لایه‌های مختلفی تشکیل‌شده است که هر لایه از آن باید بر اساس نیاز و متناسب با کارکرد مورد انتظار شبکه انتخاب گردد. البته این انتخاب فقط مربوط به لایه‌ها نمی‌گردد بلکه تعداد سلول‌های[۱۰] مورداستفاده نیز باید هوشمندانه به‌طوری انتخاب گردد که بازدهی سیستم در حداکثر و حجم محاسبات در حداقل قرار گیرد. پس‌ازاین مورد، تنظیمات کلی شبکه عصبی اعم از InitialLearnRate ,MiniBatchSize ,ValidationData Shuffle ,GradientThreshold ,ValidationFrequency ,Plots ,Verbose انجام می‌شود. تنظیمات نحوه کلی کارکرد سیستم را تعیین می‌نماید و در دقت و بازده کلی تأثیر به سزایی دارد. پس از طراحی و پیاده شده شبکه عصبی مصنوعی، این شبکه بر اساس داده‌هایی که قبلاً پیش‌پردازش شده

[1] Vanishing Gradient.
[2] Fully Connected.
[3] Skip Connections.
[4] Residual Connections.
[5] Recurrent Neural Network.
[6] Sequential Data.
[7] Gate.
[8] Dataset.
[9] Normalization.
[10] Neuron.

است در فرآیند یادگیری قرار می‌گیرد در طی این فرآیند شبکه عصبی مصنوعی همچون ذهن انسان شروع به یادگیری وظیفه محول شده کرده و سعی در برقراری ارتباط بین داده‌های ورودی و برچسب متناظر با آن‌ها در خروجی می‌کند. پس از اتمام فرآیند یادگیری، شبکه گزارش‌هایی را در رابطه با میزان بازده و یادگیری خود اعلام می‌کند که در صورت مناسب و کارا بودن این نتایج می‌توان گفت فرآیند یادگیری به پایان رسیده است و شبکه قابل بهره‌برداری در وظیفه‌ی خود است. در انتها برای اثبات میزان دقت و کارایی شبکه می‌توان آن را بار دیگر و توسط مجموعه دادگان متفاوت دیگر مورد آزمایش قرار دارد باید توجه داشت این مرحله اختیاری بوده و برای اطمینان هر چه بیشتر انجام می‌گیرد. پس‌ازاین مرحله شبکه آماده استفاده و بهره‌برداری است و می‌توان از آن استفاده نمود. شکل (۲) مراحل طراحی و پیاده‌سازی سیستم‌های هوشمند تشخیص خشونت را به‌صورت بلوک دیاگرام متنی و گرافیکی نمایش می‌دهد.

شکل(۲): بلوک دیاگرام پژوهش



۲-۵. انتخاب مجموعه داده

در این پژوهش باتوجه‌به ماهیت تصویری آن، مجموعه داده انتخابی بر پایه مجموعه‌ای از فیلم‌های دوربین‌های مداربسته است که به دو گروه فیلم‌های خشونتی و غیر خشونتی تقسیم می‌گردد. مجموعه داده انتخابی باید ویژگی‌های زیر را داشته باشد:

۱. جامع و کامل؛

۲. کیفیت مناسب جهت تشخیص فریم‌ها؛

۳. مدت‌زمان مناسب برای هر فیلم؛

۴. فیلم با صحنه‌های مختلف خشونتی و غیر خشونت؛

۵. تعداد مناسب و زیاد جهت یادگیری بیشتر.

باتوجه‌به ویژگی‌های گفته‌شده در فوق و همچنین بررسی‌های میدانی بسیار دو مجموعه داده RWF-2000 و Survelliance fight انتخاب گردید.

مجموعه داده RWF-2000 شامل ۲۰۰۰ ویدیو است که از این مقدار ۱۰۰۰ عدد مربوط به فیلم‌های خشونت‌آمیز و ۱۰۰۰ عدد نیز مربوط به تصاویر غیر خشونتی می‌گردد. این فیلم‌ها عمدتاً توسط دوربین‌های مداربسته و درصحنه‌های واقعی جمع‌آوری‌شده‌اند و سناریوهای مختلف خشونت اعم از نزاع و درگیری گروهی و فردی را شامل می‌شوند این مجموعه داده یکی از کامل‌ترین و جامع‌ترین و دقیق‌ترین مجموعه داده‌های موجود در این حوزه است و همواره مورداستفاده بسیاری پژوهش‌ها قرارگرفته است و از آن می‌توان به‌عنوان مهم‌ترین مجموعه داده برای پژوهش تشخیص خشونت نام برد.

یکی دیگر از مجموعه داده‌های استفاده‌شده در این تحقیق Survelliance fight است این مجموعه داده نیز شامل ۳۰۰ ویدئو خشونتی و غیر خشونتی است که تماماً از نزاع و درگیری‌های واقعی ضبط‌شده توسط دوربین‌های مداربسته به‌دست‌آمده است و شامل سناریوهای مختلف دوربین‌های مداربسته است این مجموعه داده نیز یکی از کاراترین و دقیق مجموعه داده‌های موجود در این حوزه است. باتوجه‌به توضیحات فوق از مجموعه داده RWF-2000 برای آموزش سیستم هوشمند و شبکه عصبی استفاده‌شده است و از مجموعه داده Survelliance Fight نیز برای آزمون و درک بهتر کارایی شبکه در قسمت نتایج استفاده‌شده است. در انتها باید اشاره کرد که مجموع هر دو مجموعه داده استفاده‌شده در پژوهش حجمی برابر ۱۴ گیگابایت دارد و در دو فرمت avi. و mp4. ارائه‌شده است.

۲-۶. پیش‌پردازش داده‌ها

در این مرحله فیلم‌های موجود در مجموعه داده به فریم‌های متوالی

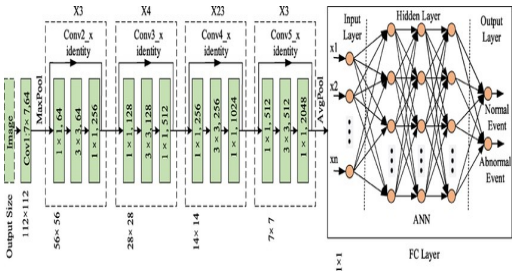
تقسیم می‌گردند. این فریم‌ها برخلاف تصاویر دارای توالی زمانی نیز است؛ لذا این توالی نباید جابه‌جا گردد؛ زیرا با جابه‌جایی فریم‌های مختلف تصویر شبکه قادر به تشخیص خشونت در تصاویر نخواهد. پس از عملیات جداسازی، تمامی فریم‌های تصویر به ماتریس‌های چهاروجهی شامل طول، عرض، شماره ویدئو و شماره فریم متناظر ویدئو تقسیم می‌گردد. پس‌ازاین تقسیم‌بندی تمامی تصاویر به ابعاد ۳*۲۲۴*۲۲۴ تبدیل می‌گردد این ابعاد برای ورود به شبکه عصبی ResNet101 موردنیاز است.

۲-۷. تشکیل زوج ورودی-خروجی

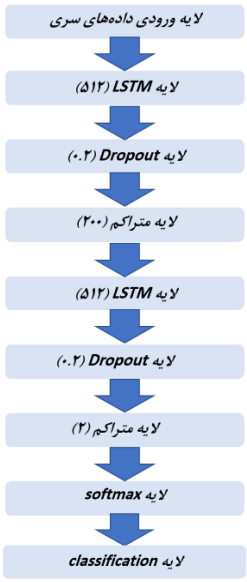
در این مرحله مجموعه تصاویر استخراج‌شده از فریم‌های متوالی تصویر برچسب‌گذاری خواهد شد باتوجه‌به ماهیت این پژوهش که تشخیص خشونتی و یا غیر خشونتی بودن تصاویر دوربین‌های مداربسته است مسئله به‌صورت کلاسه‌بندی تعریف می‌شود و در دو کلاس خشونت و غیر خشونت قرار می‌گیرد؛ لذا تمامی فریم‌های اصلاح‌شده مرحله قبل بر اساس محتوی تصویر به یکی از دو عنوان تصاویر شمال خشونت و یا تصاویر فاقد خشونت تقسیم‌بندی می‌گردد. این مرحله در این پژوهش با استفاده نام فولدر محتویات موجود انجام می‌گیرد بدین‌صورت که در ابتدا تمامی ویدئوهای خشونتی در فایل بانام تصاویر خشونت‌آمیز قرارگرفته و به طبع آن تمامی ویدئوهای غیر خشونتی نیز در یک فایل بانام تصاویر غیر خشونتی قرار می‌گیرد. سپس با استفاده از کدهای دستوری متلب برای هر ویدئو برچسبی بانام فایلی که در آن قرار دارد، داده می‌شود این ویدئو به‌عنوان ورودی و برچسب آن به‌عنوان خروجی شبکه است و سیستم خبره سعی دریافتن رابطه‌ای بین این دو مورد است. تا از طریق به تواند برای ویدئوهای آتی نیز کار تشخیص خشونت را انجام دهد.

۲-۸. طراحی و پیاده‌سازی شبکه عصبی

این مرحله مهم‌ترین و پیچیده‌ترین مرحله پژوهش است در این مرحله بامطالعه و بررسی دقیق بیش از ۲۰ شبکه آموزش‌دیده و همچنین با بررسی این شبکه‌ها ازنظر دقت و کارایی و حجم محاسباتی در سیستم هوشمند اصلی نتیجه‌گیری شد که شبکه ResNet101 می‌تواند علاوه بر حجم کمتر محاسباتی نسبت به دیگر شبکه‌ها، دقت بالاتری را داشته باشد. در این مجموعه، شبکه عصبی ResNet101 وظیفه‌ی استخراج ویژگی‌های مکانی و زمانی را از فریم‌های مختلف تصویر دارد. این ویژگی‌ها در ماتریس‌هایی با ابعاد ۲۲*۱۰۲۴ ذخیره می‌گردد که عدد ۱۰۲۴ مربوط به تعداد ویژگی‌های استخراج‌شده از شبکه ResNet101 است و عدد ۲۲ نیز تعداد فریم‌های موجود در هر ماتریس را نمایش می‌دهد. مجموع تمامی این ماتریس‌ها برابر ۱۶۰۰ عدد است به عبارتی شبکه ResNet101 با ۴۴,۶ میلیون پارامتر در داخل خود می‌تواند ۷۲۰۸۶۶۰۰ ویژگی از کل مجموعه داده استخراج نماید. در ادامه با توجه به تعداد خروجی شبکه ResNet101 که ماتریسی با ابعاد ۲۲*۱۰۲۴ است بهینه‌ترین مقدار برای تعداد نورون‌ها در شبکه حافظه‌دار LSTM، ۱۰۲۴ عدد نورون است ولی باید توجه داشت که به دلیل پیچیدگی‌های بسیار در ویژگی‌های فضا-زمانی داده‌های فریم‌های متوالی تصویر استفاده از تمامی نورون‌ها در یک‌لایه دقت مناسبی را ارائه نمی‌نماید لذا در این پژوهش از دولایه Lstm در کنار دولایه dropout ((برای جلوگیری از برازش بیش‌ازحد)) استفاده شد. برای درک هر چه‌بهتر و مناسب‌تر شبکه زمانی از دولایه متراکم نیز پس از لایه‌های dropout استفاده گردیده است که این دولایه در درک هرچه بهتر ویژگی‌های تصویر کمک شایانی به سیستم کلی نموده است. در انتها شبکه حافظه‌دار نیز یک‌لایه softmax و classification برای ایجاد خروجی موردنظر قرار داده‌شده است که از این طریق بتواند خروجی مطلوب که برچسب خشونتی و یا غیر خشونتی فیلم‌ها است را ایجاد نماید. شکل ۳ نمای کلی از شبکه ResNet101 و شکل ۴نمایی از شبکه LSTM طراحی‌شده در این پژوهش را به‌صورت مجزا از یکدیگر نمایش می‌دهد.



شکل(۳): معماری شبکه ResNet101 [۱۴]



شکل(۴): معماری شبکه ResNet101

در ادامه روند کلی توسعه و پیاده‌سازی شبکه عصبی شرح که در نرم‌افزار متلب پیاده‌سازی شده است شرح داده می‌شود.

شبیه‌سازی و پیاده‌سازی سیستم خبره تشخیص خشونت در

دوربین‌های مداربسته از سه قسمت عمده زیر تشکیل‌شده است:

۱. تبدیل ویدئوها به ویژگی‌های دنباله‌دار زمانی با استفاده از یک شبکه عصبی پیچشی از پیش آموزش‌دیده، مانند ResNet101 ؛

۲. آموزش یک شبکه عصبی حافظه‌دار به‌مانند شبکه عصبی LSTM برای تشخیص خشونت با استفاده از ویژگی‌های زمانی استخراج‌شده در مرحله قبل؛

۳. آموزش هر دو مجموعه برای انجام وظایف تعریف‌شده؛

۴. اتصال شبکه عصبی پیچشی به شبکه عصبی LSTM برای تشکیل یک سیستم یکپارچه خبره.

شکل (5) نمای کلی شبکه طراحی‌شده در نرم‌افزار متلب را که در آن دو شبکه استخراج گر و زمانی با یکدیگر ترکیب‌شده‌اند نمایش داده‌شده است.

شکل(5). نمایی از شبکه طراحی‌شده

همان‌گونه که در شکل فوق مشاهده می‌شود برای ورودی شبکه از یک‌لایه با عنوان لایه ورودی داده‌های سری استفاده‌شده است. این لایه برای دریافت داده‌های ورودی فریم‌های مختلف تصویر استفاده‌شده است. برای استخراج ویژگی‌های تصویر، یعنی اعمال عملیات پیچش روی هر فریم از فیلم‌ها به‌طور مستقل، از یک‌لایه sequence Folding و به دنبال آن لایه‌های پیچشی استفاده‌شده است تا ویژگی‌های مکانی-زمانی موجود در فریم‌های مختلف تصویر استخراج گردد. در ادامه برای بازیابی ساختار توالی تصاویر و تغییر شکل خروجی به دنباله‌های برداری متناظر با آن، از یک‌لایه sequence Unfolding و یک‌لایه Flatten استفاده‌شده است. در آخر نیز از یک شبکه عصبی حافظه‌دار بر پایه LSTM برای شناسایی و تشخیص الگوهای استخراج‌شده و برچسب‌گذاری بر روی‌داده‌های جدید استفاده‌شده است.

۳. نتایج و بحث

برای تشخیص میزان دقت در شبکه‌های عصبی مصنوعی از روابط مختلفی می‌توان استفاده نمود؛ ولی باید توجه داشت که این معیارها باتوجه‌به ماهیت مسئله یعنی کلاسه‌بندی یا رگرسیونی بودن مسئله می‌تواند با یکدیگر متفاوت باشد ازجمله معیارهای متداول استفاده در مسائل رگرسیونی می‌توان به MSE[1], MAPE[2], RMS[3] و MAE[4] اشاره داشت که مقدار خطای سیستم را بر اساس عددی نمایش می‌دهد در مسائل مربوط به کلاسه‌بندی نیز باتوجه‌به ماهیت مسئله که نگاشت مجموعه ورودی بر روی کلاسی از مجموعه خروجی است، استفاده از روش‌های مانند ماتریس درهم‌ریختگی[5] توصیه می‌گردد. ماتریس درهم‌ریختگی به‌خوبی از طریق تعداد نمونه‌های صحیح و غلط می‌تواند میزان بازدهی و کارایی شبکه عصبی و سیستم خبره را تشخیص دهد. از روی ماتریس درهم‌ریختگی می‌توان به‌صورت عددی و از طریق رابطه زیر میزان دقت برای شبکه عصبی را به دست آورد. معادله ۱ میزان دقت سیستم را بر اساس ماتریس درهم‌ریختگی نشان می‌دهد.

$$accurecy = \frac{TN+TP}{TN+TP+FN+FP} * 100 \qquad (1)$$

در معادله فوق موارد صحیح در صورت کسر و تمامی موارد اعم از صحیح و غلط دیر کسر قرار می‌گیرد. یکی دیگر از معیارهای بسیار مناسب برای تشخیص میزان دقت شبکه مربوط به تابع خطا است تابع خطا در این پژوهش بر اساس یکی از روش‌های متداول اندازه‌گیری میزان مطابق در مسائل کلاسه‌بندی یعنی تابع آنتروپی متقاطع[6] است. درواقع تابع آنتروپی متقاطع میزان تفاوت بین دو توزیع احتمال را مشخص می‌کند. بدین‌صورت که تابع softmax موجود در لایه یکی مانده به آخر شبکه عصبی خروجی بر اساس میزان احتمال تعلق بردار ورودی به هر یک از کلاس‌های تعریف‌شده در خروجی را محاسبه می‌کند و سپس این مقدار احتمال توسط توابع لگاریتمی با مقدار احتمال واقعی مقایسه می‌گردد و مقداری را ایجاد می‌نماید. این مقدار میزان تفاوت را نشان می‌دهد به عبارتی بهتر چنانچه تابع آنتروپی متقاطع شبکه به صفر میل کند نشان از شباهت زیاد مقادیر کلاس‌های پیش‌بینی‌شده باکلاس‌های واقعی را دارد و نشانی از کارایی و دقیق‌تر بودن شبکه دارد. به‌عبارتی‌دیگر خروجی واحد فعال‌سازی softmax در شبکه عصبی به‌صورت رابطه ۲ است.

$$y_r(x) = \frac{exp(a_r(x))}{\sum_{j=0}^{k} exp(a_j(x))} \qquad (2)$$

در رابطه فوق $0 \le y_r \le 1$ ، $\sum_{j=1}^{k} y_j = 1$ برقرار است. فعال‌سازی لایه‌های softmax را به‌صورت رابطه ۳ نیز می‌توان تعریف کرد.

$$p(c_r|x,\theta) = \frac{p(x,\theta|c_r)p(c_r)}{\sum_{j=1}^{k} p(x,\theta|c_j)p(c_j)} \qquad (3)$$

در رابطه فوق $0 \le p(c_r|x,\theta) \le 1$ ، $\sum_{j=1}^{k} p(c_j|x,\theta) = 1$ برقرار است و $p(c_r|x,\theta)$ نشان‌دهنده احتمالی شرطی مربوط به کلاس مربوطه در شبکه عصبی است.

برای شبکه‌های عصبی طبقه‌بندی لایه طبقه‌بندی معمولاً از یک‌لایه softmax پیروی می‌کند. در لایه طبقه‌بندی، تابع trainNetwork مقادیر تابع softmax را می‌گیرد و هر ورودی را با استفاده از تابع آنتروپی متقاطع به یک کلاس مرتبط

[1] Mean square error
[2] Mean absolute percentage error
[3] Root mean squar error
[4] Mean absolute error

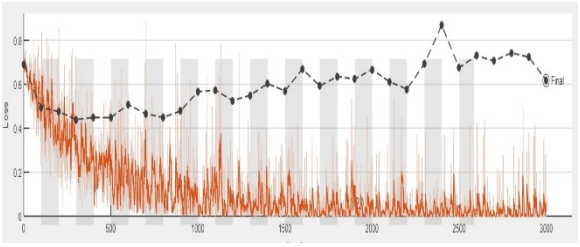
[5] Confusion matrix
[6] Cross Entropy Loss

می‌سازد تابع خطای آنتروپی از رابطه ۴ برای محاسبه میزان خطا که به‌عنوان یکی از پارامترهای مشخص‌کننده توان شبکه عصبی در تعیین کلاس‌های درست است، استفاده می‌نمایند.

$$loss = -\frac{1}{N}\sum_{n=1}^{N}\sum_{i=1}^{K} w_i t_{ni} ln y_{ni} \qquad (4)$$

در رابطه فوق N تعداد نمونه‌های موجود، K تعداد کلاس‌های موجود، W وزن کلاس i ، t نشانگر تعلق نمونه i ام به کلاس n ام و y احتمال است که شبکه برای تعلق نمونه i ام به کلاس n ام در نظر می‌گیرد. حال پس از توضیحاتی مختصر دررابطه‌با معیارهای راستی آزمایی سیستم خبره به تعیین میزان دقت و کارایی شبکه ارائه‌شده در این پژوهش پرداخته خواهد شد.

سیستم خبره برای هر دو مجموعه داده اجرا شد. در این مجموعه داده‌ها همان‌طور که پیش‌تر نیز ذکر شد درمجموع ۱۹۰۰ ویدئو با نسبت مساوی خشونتی و غیر خشونتی از دوربین‌های مداربسته موجود است باید توجه داشت که این ویدئوها از تصاویر خشونتی واقعی تشکیل‌شده است و ساختگی نیست؛ لذا شبکه از همان ابتدا بر روی‌داده‌هایی باکیفیت پایین و واقعی آموزش‌دیده است که این مهم تأثیر بسیار زیادی بر روی کارایی و راندمان این شبکه دارد و می‌توان گفت این سیستم خبره برای کاربر روی ویدئوهایی باکیفیت پایین و نویز دار بهینه‌سازی است. از مجموع ۱۹۰۰ ویدئو ۷۵ درصد آن برای آموزش و ۲۵ درصد آن برای اعتبارسنجی استفاده گردید. شکل 6 نمایی کلی از نتیجه حاصل



بر اساس دقت را در طول کل فرآیند آموزش شبکه نشان می‌دهد.

شکل(۶). نمایی از دقت شبکه

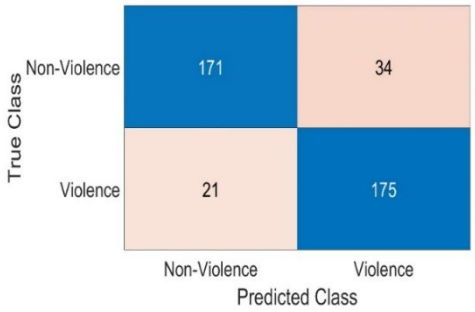
لازم به توضیح است که این شبکه بعد از هر ۱۰۰ تکرار یکبار میزان دقت را بر روی‌داده‌های اعتبارسنجی آزمایش می‌نمایند که نقاط توپر سیاه‌رنگ بر روی شکل نیز نشان‌دهنده دقت داده‌های اعتبارسنجی است. نمودار با خط ممتد نیز در شکل فوق مربوط به‌دقت شبکه در حین یادگیری است. شکل 7 نمودار تابع آنتروپی متقابل را برای این شبکه نمایش می‌دهد. در شکل 7 نیز خط با دایره توپر مربوط به نتایج حاصل از اعتبارسنجی داده‌ها و خط ممتد نیز مربوط به مقدار آنتروپی در طول فرآیند آموزشی است.

شکل(7). نمایی از دقت شبکه بر اساس آنتروپی

همان‌گونه که در شکل مشاهده می‌شود و در فوق نیز توضیح داده شد. تابع خطا نشانگر میزان مطابقت خروجی پیش‌بینی‌شده و

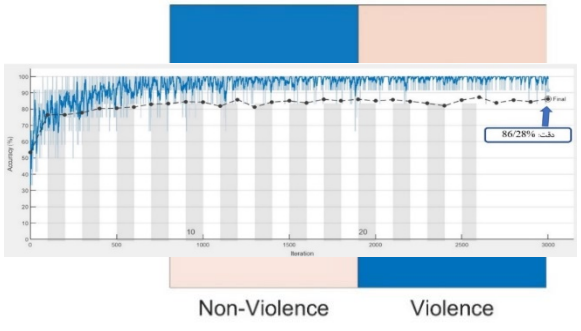
خروجی واقعی است و هرچه مقدار آن کمتر باشد نشان از درک صحیح شبکه نسبت به وظیفه داده‌شده به آن را دارد. همان‌گونه که در شکل ۷ دیده می‌شود این میزان خطا در شبکه ارائه‌شده در این تحقیق برابر ۰٫۶ است که باتوجه‌به مقدار کم عدد می‌توان نتیجه‌گیری کرد که شبکه توانسته است وظیفه تشخیص خشونت در دوربین‌های مداربسته را با دقت و کارایی مناسب تشخیص دهد. با توجه داشت که دو شکل ۶ و ۷ مکمل یکدیگر است و هر دو به‌صورت برخط و آنلاین توسط تولباکس یادگیری عمیق نرم‌افزار متلب رسم شده‌اند.

در ادامه برای درک بهتر کارایی سیستم از ماتریس درهم‌ریختگی استفاده شد. این ماتریس دارای تعاریف کیفی مناسبی است که می‌تواند به درک دقیق‌تر خوانندگان نسبت به‌دقت سیستم کمک شایانی نماید در شکل ۸ ماتریس درهم‌ریختگی برای داده‌های اعتبارسنجی شبکه بر روی‌داده‌های مجموعه داده RWF2000 مشاهده می‌شود. در انتها نیز برای اعتبارسنجی مناسب‌تر سیستم هوشمند از مجموعه داده Surveillance fight استفاده شد.



شکل (۸): ماتریس درهم‌ریختگی نتایج

شکل 9 نتایج حاصل از این مجموعه داده را بر روی سیستم طراحی‌شده پژوهش نشان می‌دهد. باید توجه داشت که این ماتریس بر روی‌داده‌های آموزش مورداستفاده قرار نمی‌گیرد بلکه از آن بر روی‌داده‌های اعتبارسنجی که از همان ابتدای کار برای بررسی میزان دقت شبکه کنار گذاشته‌شده است، استفاده می‌شود لذا می‌توان ماتریس درهم‌ریختگی را به‌عنوان کارایی کیفی سیستم خبره در نظر گرفت.



شکل (۹): ماتریس درهم‌ریختگی نتایج برای داده‌های آزمون

همان‌گونه که از شکل فوق مشاهده می‌شود برای قسمت اعتبارسنجی شبکه درمجموع از ۴۰۱ ویدئو استفاده‌شده است. شبکه ۱۷۱ ویدئوی خشونتی را به‌درستی خشونتی تشخیص داده و گزارش نموده است.

همچنین ۱۷۵ ویدئو را که غیر خشونتی بوده به‌درستی غیر خشونتی تشخیص داده است که این دو مورد نشان از میزان دقت بالای شبکه در تشخیص خشونت است. از طرفی باتوجه‌به اینکه شبکه دارای دقت ۸۶ درصدی است لذا ۲۱ مورد ویدئو خشونتی را شبکه نتوانسته است به‌درستی تشخیص دهد. در ادامه برخی دیگر از معیارهای سنجش و کارایی شبکه به‌مانند پوشش[1] و صحت[2] و امتیاز F[3] به همراه روابط حاکم بر آن‌ها و مقادیر متناظر با آن در این پژوهش آورده شده است.

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1Score = \frac{2*(Percision*Recall)}{Percision+Recall} \quad (6)$$

$$MCC = \frac{TP*TN-FP*FN}{\sqrt{(TP+FP)(TP+FP)(TN+FP)(TN+FN)}} \quad (7)$$

جدول(۱). معیارهای سنجش کارایی شبکه

معیار	نتیجه
Accuracy *(دقت)*	۸۶/۲۸٪
Recall (پوشش)	0.8929
Percision (صحت)	0.8373
F1score (امتیاز f)	۰,۸۶۴۲
MCC	۰,۷۲۷۵

همان‌گونه که از جدول فوق مشاهده می‌شود شبکه با دقت و کارایی بسیار مناسبی توانسته است به تشخیص خشونت در دوربین‌های مداربسته بپردازد.

در ادامه برای انجام بررسی‌های بهتر و جامع‌تر از نتایج حاصل از این پژوهش، مقایسه‌ای با پژوهش‌های دیگر انجام‌شده است البته باید توجه داشت که امکان بررسی موشکافی در این پژوهش‌ها وجود ندارد. زیرا در نتایج حاصل از این تحقیق فاکتورهایی مانند کاهش حجم محاسباتی در کنار حفظ میزان دقت به همراه عملکرد بلادرنگ و برخط مدنظر بوده است و سیستم خبره به گونه طراحی‌شده است که در برابر کیفیت پایین ویدئوها، وجود نویز در داده‌ها و زمان اندک ویدئو مقاوم باشد و بتواند دقت مناسب خود را حفظ نماید. این مهم در بیشتر مقالات و پژوهش‌های پیشین معیار نبوده و درنتیجه رعایت نشده است. در جدول ۲ نتایج حاصل از پژوهش‌های پیشین و این پژوهش به همراه نقاط قوت و ضعف آن آورده شده است.

دقت(٪)	روش	نام نویسندگان
۸۶/۲۸	شبکه عصبی ResNet101 و LSTM	روش این پژوهش
۸۰	تشخیص ویژگی‌های فضا-زمان مانند شتاب و...	گارسیا[1]
۷۹	شبکه convNet با روش یادگیری انتقال	دینگ [10]
۸۲	ResNet18بر روی‌داده‌های صوتی	سانتوس[11]
۸۳/۳۴	تئوری bag-of-words	فبین [5]
۹۰	استخراج ویژگی و طبقه‌بندی SVM	چادهاری [9]
۸۷/۲۵	3D-CNN و جریان نوری	پراتاما [6]

همان‌گونه که در جدول ۲ مشاهده می‌شود سیستم خبره ارائه‌شده در این پژوهش با دقت مناسبی می‌تواند به تشخیص خشونت در دوربین‌های مداربسته بپردازد. باید توجه داشت دو پژوهش چادهاری و پراتاما اندکی دقت بالاتر نسبت به شبکه ارائه‌شده در این مقاله دارند زیرا درروش چادهاری و همکاران از مجموعه داده ویدیوهای خشونت در فیلم‌های سینمایی استفاده‌شده است که علاوه بر کیفیت بسیار بالاتر نسبت به مجموعه داده استفاده‌شده در این پژوهش، دارای تصاویری ساده‌تر و بدیهی‌تر است که کار تشخیص خشونت برای سیستم هوشمند را ساده‌تر می‌نماید. در رابطه با پژوهش پراتاما و همکاران نیز باید اشاره کرد که روش ایشان بلادرنگ نبوده و از حجم بالایی از محاسبات برخوردار است به‌عنوان نمونه در این پژوهش ۳۰ فریم برای هر ثانیه از ویدیو استفاده‌شده است. درصورتی‌که روش این مقاله فقط از ۲۲ فریم به ازای هر ۵ ثانیه ویدیو استفاده کرده است. در انتها می‌توان اشاره داشت که روش ارائه‌شده در این پژوهش علاوه بر حفظ دقت توانسته است حجم محاسباتی بسیار کمتری نسبت به پژوهش‌های پیشین داشته باشد. همچنین باید توجه داشت که مجموعه داده مورداستفاده نیز از کیفیت پایین ۲۲۴*۲۲۴ پیکسلی برخوردار است ک نشان از مقاومت بالای سیستم هوشمند در برابر وجود نویز در تصاویر دارد.

۱,۳ بحث و پیشنهادها

در این پژوهش سیستمی هوشمند بر اساس فناوری بینایی ماشین و شبکه‌های عصبی مصنوعی پیچشی و حافظه‌دار ارائه شد. این سامانه هوشمند می‌تواند با دقت بسیار مناسب و در کمترین زمان ممکن به تشخیص خشونت در دوربین‌های مداربسته بپردازد و در صورت بروز هرگونه خشونت مراتب را به مأمورین ذی‌ربط اطلاع دهد. به‌عنوان پیشنهادهایی برای پژوهش‌های آینده توصیه می‌گردد که از روش‌های یادگیری عمیق تقویتی که ازجمله جدیدترین روش‌های آموزشی است استفاده گردد تا میزان دقت شبکه افزایش یابد. همچنین برای هر چه گستردگی و جامعیت پژوهش‌های آینده پیشنهاد می‌گردد تشخیص خشونت به همراه نوع آن انجام گیرد. به‌عنوان‌مثال سامانه هوشمند مجهز به ۱۰ زیرمجموعه برای خشونت اعم از نزاع، چاقوکشی، سلاح،

[1] Rrcall
[2] Percision
[3] F1-Score

دزدی، تجاوز و ... گردد کـه عـلاوه بـر تشـخیص خشـونت بتوانـد نـوع خشونت موجود در سامانه را نیز تشخیص دهد.

۴. مراجع

[1] I. Serrano Gracia, O. Deniz Suarez, G. Bueno Garcia, and T.-K. Kim, "Fast fight detection," *PloS one,* vol. 10, no. 4, p. e0120448, 2015.

[2] S. U. Khan, I. U. Haq, S. Rho, S. W. Baik, and M. Y. Lee, "Cover the violence: A novel Deep-Learning-Based approach towards violence-detection in movies," *Applied Sciences,* vol. 9, no. 22, p. 4963, 2019.

[3] P. Wu, J. Liu, Y. Shi, Y. Sun, F. Shao, Z. Wu, and Z. Yang, "Not only look, but also listen: Learning multimodal violence detection under weak supervision," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, 2020: Springer, pp. 322-339.

[4] J. Mahmoodi and A. Salajeghe, "A classification method based on optical flow for violence detection," *Expert systems with applications,* vol. 127, pp. 121-127, 2019.

[5] I. Febin, K. Jayasree, and P. T. Joy, "Violence detection in videos for an intelligent surveillance system using MoBSIFT and movement filtering algorithm," *Pattern Analysis and Applications,* vol. 23, no. 2, pp. 611-623, 2020.

[6] R. A. Pratama, N. Yudistira, and F. A. Bachtiar, "Violence recognition on videos using two-stream 3D CNN with custom spatiotemporal crop," *Multimedia Tools and Applications,* pp. 1-23, 2023.

[7] A. Traoré and M. A. Akhloufi, "2D bidirectional gated recurrent unit convolutional neural networks for end-to-end violence detection in videos," in *International Conference on Image Analysis and Recognition*, 2020: Springer, pp. 152-160.

[8] A. Srivastava, T. Badal, and R. Singh, "Real life violence detection in surveillance videos using spatiotemporal features," in *2021 Thirteenth International Conference on Contemporary Computing (IC3-2021)*, 2021, pp. 262-266.

[9] S. Chaudhary, M. A. Khan, and C. Bhatnagar, "Multiple anomalous activity detection in videos," *Procedia Computer Science,* vol. 125, pp. 336-345, 2018.

[10] C. Ding, S. Fan, M. Zhu, W. Feng, and B. Jia, "Violence detection in video by using 3D convolutional neural networks," in *Advances in Visual Computing: 10th International Symposium, ISVC 2014, Las Vegas, NV, USA, December 8-10, 2014, Proceedings, Part II 10*, 2014: Springer, pp. 551-558.

[11] F. Santos, D. Durães, F. S. Marcondes, N. Hammerschmidt, S. Lange, J. Machado, and P. Novais, "In-car violence detection based on the audio signal," in *International conference on intelligent data engineering and automated learning*, 2021: Springer, pp. 437-445.

[12] T. Zhang, Z. Yang, W. Jia, B. Yang, J. Yang, and X. He, "A new method for violence detection in surveillance scenes," *Multimedia Tools and Applications,* vol. 75, pp. 7327-7349, 2016.

[13] A. Mumtaz, A. B. Sargano, and Z. Habib, "Violence detection in surveillance videos with deep network using transfer learning," in *2018 2nd European Conference on Electrical Engineering and Computer Science (EECS)*, 2018: IEEE, pp. 558-563.

[14] Torabipour, Tobi, Siadat, Sayeda Safiya. A method for predicting the stock price of Tehran stock market in relation to knowledge. electronic and cyber defense, 1401; 10 (4): 91-100[in persian]

[15] F. J. Rendón-Segador, J. A. Álvarez-García, F. Enríquez, and O. Deniz, "Violencenet: Dense multi-head self-attention with bidirectional convolutional lstm for detecting violence," *Electronics,* vol. 10, no. 13, p. 1601, 2021.

[16] B. M. Peixoto, S. Avila, Z. Dias, and A. Rocha, "Breaking down violence: A deep-learning strategy to model and classify violence in videos," in *Proceedings of the 13th International Conference on Availability, Reliability and Security*, 2018, pp. 1-7.

[17] A. Jain and D. K. Vishwakarma, "Deep NeuralNet for violence detection using motion features from dynamic images," in *2020 third international conference on smart systems and inventive technology (ICSSIT)*, 2020: IEEE, pp. 826-831.

[18] Z. Islam, M. Rukonuzzaman, R. Ahmed, M. H. Kabir, and M. Farazi, "Efficient two-stream network for violence detection using separable convolutional lstm," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021: IEEE, pp. 1-8.

[19] S. Vosta and K.-C. Yow, "A cnn-rnn combined structure for real-world violence detection in surveillance cameras," *Applied Sciences,* vol. 12, no. 3, p. 1021, 2022.

[20] Irfanullah, T. Hussain, A. Iqbal, B. Yang, and A. Hussain, "Real time violence detection in surveillance videos using Convolutional Neural Networks," *Multimedia Tools and Applications,* vol. 81, no. 26, pp. 38151-38173, 2022.

[21]. V. D. Huszár, V. K. Adhikarla, I. Négyesi, and C. Krasznay, "Toward Fast and Accurate Violence Detection for Automated Video Surveillance Applications," *IEEE Access,* vol. 11, pp. 18772-18793, 2023.

[22] S. Singh and B. Tyagi, "Computational Comparison of CNN Based Methods for Violence Detection," 2023.

[23] A. L. Detzel, H. Liu, J. Strauss, G. Zhou, and Y. Zhu, "Bitcoin: Learning and predictability via technical analysis," in *Paris December 2018 Finance Meeting EUROFIDAI-AFFI*, 2019, vol. 1, p. 314.

[24] T. Zhang, Z. Yang, W. Jia, B. Yang, J. Yang, and X. He, "A new method for violence detection in surveillance scenes," *Multimedia Tools and Applications,* vol. 75, pp. 7327-7349, 2016.

[25] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Physica D: Nonlinear Phenomena,* vol. 404, p. 132306, 2020.