RCALAD: Regularized Complete Adversarially Learned Anomaly detection

چکیده

یکی از مهم‌ترین فعالیت‌های حوزه تحلیل داده تشـــخیص ناهنجاری می‌باشـــد که در طیف وســـیعی از کاربردها همچون تشـــخیص جعل، کاربردهای پزشکی و سیستم‌های امنیتی به کار گرفته می‌شود. علی‌رغم وجود روش‌های آماری و مبتنی بر یادگیری ماشین طراحی مدل‌های موثر در تشخیص ناهنجاری در فضای داده پیچیده با ابعاد بالا همچنان به عنوان یک چالش اساسی باقی مانده است. شبکه‌های مولد تخاصمی قادرند تا بر چالش مورد نظر فائق آمده و توزیع داده‌های دنیای واقعی که دارای پیچیدگی و ابعاد بالا هســتند را مدل کنند و همین امر ســبب می‌شود تا عملکرد امیدوارکننده‌ای در زمینه تشخیص ناهنجاری از خود نشان دهند. در این کار چارچوب تخاصم RCALAD با هدف تشخیص ناهنجاری ارائه شده است. اساس کار مدل پیشنهادی بازسازی داده ورودی با استفاده از شبکه مولد و در ادامه محاسبه میزان اختلاف داده اصلی و بازسازی آن به منظور شناسایی نمونه‌های ناهنجار است. لازمه شناسایی موثر نمونه‌های ناهنجار بازسازی ضعیف داده‌های ناهنجار است. مدل پیشنهادی نقاط ضعف کارهای پیشین را پوشش داده و بر دو جنبه تضمین بازسازی ضعیف نمونه‌های ناهنجار و همچنین استفاده از بیشینه اطلاعات موجود در شبکه به صورت توامان برای آموزش بهتر، تمرکز کرده است. تشخیص ناهنجاری با استفاده از خطای بازسازی نیازمند تعریف امتیاز ناهنجاری مناسب است، بنابراین علاوه بر معماری پیشنهادی دو امتیاز ناهنجاری جدید نیز در این کار ارائه شده است. نتایج تجربی بیانگر برتری و قدرت مدل RCALAD در مقایسه با سایر مدل‌های به روز و مطرح در زمینه تشخیص ناهنجاری بوده است.

کلید واژه- تشخیص ناهنجاری، یادگیری ماشین، شبکه مولد تخاصمی، خطای بازسازی، امتیاز ناهنجاری.

۱- مقدمه

هنگام تجزیه و تحلیل دادگان موجود در دنیای واقعی، شناسایی نمونه‌های غیرمشابه با سایر نمونه‌ها امری ضروری به نظر می‌رسد. چنین نمونه‌هایی با عنوان ناهنجاری شناخته می‌شوند و از عملیات شناسایی چنین نمونه‌هایی با عنوان مسئله تشخیص ناهنجاری یاد می‌شود. این مسئله یک بخش حائز اهمیت از زمینه تحقیقاتی داده‌کاوی است چرا که شامل کشف الگوهای جذاب و نادر در داده‌هاست[1].

ناهنجاری‌ها جزو پارامترهای مهم هر مجموعه داده‌ای در نظر گرفته می‌شوند و در دامنه وسیعی از کاربردها تاثیرگذار هستند. به عنوان مثال، الگوی غیر معمول ترافیک در یک شبکه کامپیوتری میتواند به معنای هک شدن رایانه و انتقال داده‌ها به مقصدهای غیرمجاز باشد. رفتار غیر عادی در معاملاتی که توسط کارت‌های اعتباری انجام می‌شوند می‌تواند نشانگر فعالیت‌های اقتصادی با هدف کلاهبرداری باشد[2]، و یا یک ناهنجاری در تصویر MRI ممکن است وجود تومور بدخیم را نشان دهد[3]. علی‌رغم وجود روش‌های آماری و مبتنی بر یادگیری ماشین، طراحی مدل‌های موثر در تشخیص ناهنجاری در فضای داده پیچیده با ابعاد بالا همچنان به عنوان یک چالش اساسی باقی مانده است[4].

شبکه‌های مولد تخاصمی قادرند تا بر چالش مورد نظر فائق آمده و توزیع داده‌های دنیای واقعی که دارای پیچیدگی و ابعاد بالا هستند را مدل کنند و همین امر سبب می‌شود تا عملکرد امیدوارکننده‌ای در زمینه تشخیص ناهنجاری از خود نشان دهند. در شبکه‌های مولد تخاصمی، یک مدل مولد در برابر یک مدل تمایزگر قرار می‌گیرد، مدل تمایزگر سعی می‌کند میان داده‌های واقعی و داده‌های تولیدی توسط شبکه مولد تمایز ایجاد کند. در شبکه مولد تخاصمی به طور همزمان دو مدل مولد و تمایزگر آموزش داده می‌شود. مدل مولد G توزیع داده را ضبط می‌کند و مدل تمایزگر D که احتمال این که نمونه از داده‌های تولید شده توسط G باشد را تخمین می‌زند. تابع هدف برای شبکه مولد G به حداکثر رساندن احتمال اشتباه شبکه D است. این بستر منجر به یک بازی دو نفره مانند بازی‌های بیشینه-کمینه می‌شود[5]. توانایی شبکه‌های عصبی تخاصمی در مدل کردن تصاویر طبیعی ثابت شده است[6] [7]، و بر کاربرد آن‌ها در زمینه‌های پردازش گفتار و متن [4]و تصاویر پزشکی روز به روز افزوده می‌شود [8]

در این کار روشی کارآمد و موثر مبتنی بر شبکه‌های مولد تخاصمی که به صورت خاص با هدف تشخیص ناهنجاری، طراحی شده است، پیشنهاد می‌شود. مانند بسیاری از الگوریتم‌های مبتنی بر یادگیری، در این جا دو مرحله اصلی آموزش و آزمایش وجود دارد. در قسمت آموزش همانند دیگر چهارچوب‌های تخاصمی، به نوبت بخش مولد و بخش تمایزگر را آموزش می‌دهیم تا هر دو بخش در عین تناسب به نوبت به‌روزرسانی شود. در اینجا با هدف stabilize کردن آموزش ساختار تخاصمی از توزیع توام پارامترهای موجود در شبکه استفاده شده است و به منظور اعمال نگاشت معکوس نمونه‌های ورودی به فضای نهفته، به طور توام با شبکه تمایزگر و مولد، کدگذار E آموزش داده می‌شود. در بیشتر کارهای قبلی در زمینه تشخیص ناهنجاری، بر تخمین توزیع (density estimation) داده‌های هنجار تمرکز شده است و هیچ لزومی برای بازسازی ضعیف داده های ناهنجار وجود ندارد[۸]. در مدل RCALAD با تمرکز بر هدف بازسازی هر چه ضعیف‌تر نمونه‌های ناهنجار، توزیع

جریمه $\sigma(x)$ به ساختار تخاصمی پیشنهادی اضافه شده است تا کدگذار و مولد به سمت توزیع نمونه‌های هنجار بایاس شوند. آزمایش‌ها روی طیف‌های مختلف دادگان تصویری و جدولی با ابعاد بالا صورت پذیرفته‌اند و نتایج حاصل بیانگر عملکرد بهتر مدل پیشنهادی نسبت به سایر مدل‌های پایه بوده است.

۲- کارهای مرتبط

تشخیص ناهنجاری به نام‌های شناسایی نوآوری و تشخیص داده پرت نیز شناخته می‌شود، این مسئله همانطور که در[11][10][9] بررسی شد به طور گسترده مورد مطالعه قرار گرفته است. روش‌هایی قبلی که تاکنون در این زمینه مورد استفاده قرار گرفته است به طور کلی به دو دسته representation learning و generative model تقسیم می‌شوند .

روش‌های مبتنی بر representation learning با استخراج وبژگی‌های اصلی و یا یادگیری یک نگاشت از داده‌های نرمال مسئله تشخیص ناهنجاری را حل می‌کند one-class support vector machine .مرز حاشیه‌ای حول داده نرمال را پیدا می‌کند[12]. روش IF یک روش از دسته روش‌های کلاسیک یادگیری ماشین است در این روش با ویژگی‌هایی که به صورت تصادفی انتخاب شده‌اند به ساختن درخت پرداخته شده است. امتیاز ناهنجاری در این مدل میانگین فاصله تا ریشه است[13]. Deep support vector data description به اختصار DSVDDیک ابر کره به منظور محصور ساختن بازنمایی نمونه‌های نرمال را پیدا می‌کند[14]. liu and gryllias constructed frequency . domain features using cyclic spectral analysis and used them svdd frame. This method has been proved robust against outliers and can achieve a high detection rate for bearing anomaly detection.

Odin با استفاده از مقیاس‌بندی دما و اغتشاشات روی یک شبکه عصبی از پیش آموزش دیده شده به شناسایی نمونه‌های ناهنجار روی مجموعه داده‌های تصویری می‌پردازد[15].در [16] محققان رویکرد جدیدی برای شناسایی ناهنجاری‌های تصویری با آموزش مدل روی تصاویر نرمالی که با تبدیل هندسی تغییر یافته، ارائه داده‌اند. در این مدل دسته‌بند با استفاده از آماره‌های فعالساز سافتمکس امتیاز ناهنجاری را محاسبه می‌کند.

مدل‌های مولد معمولا تلاش می‌کنند تا بازسازی داده را یاد بگیرند و با استفاده از این بازسازی نمونه‌های ناهنجار را شناسایی می‌کنند[17].به عنوان مثال خودکدگذارها توزیع داده نرمال را مدل می‌کنند و از خطای بازسازی به عنوان امتیاز ناهنجاری استفاده می‌شود[19][18]. Deep structured energy based models به اختصار DSEBM یک مدل مبتنی بر انرژی یاد میگیرد و هر نمونه را به یک امتیاز انرژی نگاشت می‌کند[20]. deep autoencoding gaussian mixture model به اختصار DAGMM با استفاده از یک کدگذار برای نمونه‌های نرمال یک توزیع گوسی مخلوط تخمین می‌زند .اخیرا از شبکه‌های عصبی تخاصمی در تشخیص ناهنجاری استفاده شده است. به عنوان مثال از این ساختار برای شناسایی ناهنجاری در تصاویر پزشکی استفاده شده است .در [8] کار نگاشت معکوس به فضای نهفته با استفاده از مکانیزم backpropagationبازگشتی صورت می‌پذیرد. در [21]در ادامه کار قبلی انجام شده است، در این کار به منظور کاهش هزینه محاسباتی نگاشت به فضای نهفته به وسیله شبکه کدگذار صورت انجام می‌گیرد. در[22] مدل ارائه شده بر اساس شبکه عصبی تقابلی دوطرفه به اختصار BiGANبنا نهاده شده است. وظیفه نگاشت معکوس از فضای داده ورودی به فضای نهفته نیز بر عهده کدگذار است. بر خلاف ساختار استاندارد GAN که در آن تمایزگر تنها تصویر واقعی و تصویر تولیدی شبکه مولد را ورودی می‌گیرد، بازنمایی این تصاویر در فضای نهفته هم به عنوان ورودی به شبکه تمایزگر داده می‌شود. مدل ALAD با استفاده از چارچوب شبکه‌های عصبی تقابلی که در [23] معرفی شد و با آموزش همزمان یک کدگذار با هدف دستیابی به نگاشت معکوس از فضای داده ورودی و همچنین اضافه کردن تمایزگر D_{zz} به ساختار تقابلی با هدف پایدارسازی روند آموزش، به تشخیص ناهنجاری پرداخته است. مدل DCAE یک مدل کلاسیک خودکدگذار است که در آن کدگذار و کدگشا دارای ساختار کانولوشنی هستند. امتیاز ناهنجاری در این مدل نرم دو خطای بازسازی است[24].

۳- پیش نیازها/Preliminaries

شبکه‌های مولد تقابلی اولین بار در سال ۲۰۱۴ توسط آقای گودفلو و همکاران ابداع شد[5]، در این شبکه‌ها زیرشبکه مولد در برابر زیرشبکه تمایزگر قرار می‌گیرد، این زیرشبکه‌ها روی مجموعه M نمونه‌ای بدون برچسب $\{x^{(i)}\}_{i=1}^{M}$ آموزش می‌بینند. زیرشبکه مولد نمونه‌های انتخابی از فضای نهفته z را به فضای داده ورودی نگاشت می‌کند. زیرشبکه تمایزگر سعی می‌کند میان داده واقعی $x^{(i)}$ و داده تولیدی توسط شبکه مولد یعنی G تمایز ایجاد کند. این دو زیرشبکه با هم در رقابت هستند، شبکه مولد G تلاش می‌کند تا توزیع داده ورودی را تقلید کند در حالی‌که شبکه تمایزگر تلاش می‌کند تا میان نمونه‌های واقعی و داده تولیدی زیرشبکه مولد تمیز دهد. در فاز آموزش شبکه مولد G و شبکه تمایزگر D به صورت متناوب با استفاده از گرادیان کاهشی و به نوبت بهینه می‌شوند.

توزیع روی داده ورودی به صورت $q(x)$ نمایش داده می‌شود و $p(z)$ به عنوان شبکه مولد در فضای نهفته Z در نظر گرفته می‌شود. آموزش شبکه GAN با پیدا کردن تمایزگر و مولدی که بتواند مسئله saddle point که به شکل $\min_G \max_D V(D,G)$ است را حل کند، انجام می‌شود.

تابع $V(D,G)$ به صورت زیر تعریف می‌شود:

$$V(D,G) = E_{x\sim q(x)}[log(D(x))] + E_{z\sim p(z)}[log(1-D(G(z)))]$$

حل این مسئله در نهایت به این نتیجه همگرا می شود که توزیع مولد با توزیع داده واقعی برابر باشد. یعنی تمایزگر بهینه سراسری در صورتی به دست خواهد آمد که اگر و تنها اگر $P_G(x)=q(x)$ باشد. منظور از P_G توزیع یادگرفته شده توسط شبکه مولد است. اثبات این قضیه در مقاله [5] آمده است. در مقاله Adversarially learned inference و یا به اختصار ALI تلاش شده تا با استفاده از کدگذار $E(x)$ توزیع توام کدگذار به صورت $q(x,z) = q(x)e(z|x)$ و توزیع شبکه مولد به صورت $p(x,z) = p(z)p(x|z)$ مدل شود[25]. در اینجا $e(z|x)$ توسط شبکه کدگذار یاد گرفته می‌شود. تابع هدف مدل ALI به صورت زیر است:

$$\min_{G,E} \max_D V(D,G) = E_{q(x,z)}[log\, D(x,E(x))] + E_{p(x,z)}[log\, (-D(G(z),z)] \quad (۳)$$

در معادله بالا D_{xz} بیانگر شبکه تمایزگر است که x و z را به عنوان ورودی می‌گیرد و مقدار خروجی آن مشخص کننده این است که با چه احتمالی ورودی‌های فعلی از توزیع $q(x,z)$ نشئت گرفته است. کدگذار، شبکه مولد و تمایزگر در حالت بهینه خود قرار می‌گیرند اگر و تنها اگر $q(x,z) = p(x,z)$. اثبات این قضیه در مقاله [25] ارائه شده است.

علیرغم اینکه توزیع‌های p و q برای ما مشخص هستند ولی در عمل و حین آموزش مدل لزوما به سمت نقطه بهینه همگرا نمی‌شود. دلیل این اتفاق در مقاله[23] به مسئله پایداری چرخه که به صورت $G(E(x)) \approx x$ تعریف می‌شود نسبت داده شده است. برای حل این مسئله چارچوب ALICE پیشنهاد داد تا تمایزگر D_{xx} به ساختار شبکه ALI اضافه شود. تابع هدف این مدل به صورت زیر است:

$$\min_{E,G} \max_{D_{xz},D_{xx}} V_{ALICE} = V_{ALI} + E_{x\sim q(x)}[log\, D_{xx}(x,x) + log\, 1 - D_{xx}(x,G(E(x)))] \quad (۴)$$

در این کار نشان داده شد که بکارگیری تمایزگر D_{xx} از نظر تئوری به بهترین بازسازی برای داده ورودی خواهیم رسید[23]. برای تثبیت آموزش در مدل پایه ALICE، در [4] توزیع‌های شرطی را با اضافه‌کردن یک تمایزگر دیگر به مدل اعمال کردند و سپس عملیات نرمال‌سازی طیفی را انجام دادند. به صورت جزئی‌تر، در این‌جا یک شبکه تمایزگر D_{zz} به مدل با هدف تضمین چرخه پایداری در فضای نهفته اضافه شده که وظیفه دارد تا متغیر در فضای نهان و بازسازی آن را تا حد امکان به یکدیگر شبیه کند. با کنار هم قرار دادن بلوک پیشنهاد شده در [4] و اجزای مدل ALICE ، در نهایت تابع هزینه مدل ALAD به صورت زیر خواهد بود.

$$min_{G,E} max_{D_{xz},D_{xx},D_{zz}} V_{ALAD} = V_{ALICE} + \mathbb{E}_{z\sim p(z)}[log(D_{zz}(z,z)) + \mathbb{E}_{z\sim p(z)}[log(-D_{zz}(z,G(E(z))))] \quad (۵)$$

در این مقاله نشان داده شده که با افزودن قیود لیپسچیتز[1] به تمایزگر شبکهGAN ، فاز آموزش تثبیت خواهد شد همچنین در عمل نشان داده شده که با spectral normalization پارامترهای وزن، روی عملکرد شبکه بهبود خواهیم داشت.

با اینکه ایده alad به پایدار شدن چرخه کمک می‌کند اما همچنان متغیرهای فضای نهفته و ورودی را در دو فضای مستقل از هم بررسی شوند و از وابستگی ذاتی میان متغیرها چشم‌پوشی می‌شود. به صورت دقیق‌تر متغیرهای x و $\hat{x}$ در یک روند به صورت جدا از روند بررسی متغیرهای z و $\hat{z}$ بررسی می‌شود. در صورتی که روند بازسازی این دو جفت داده در طول یکدیگر قرار دارند و بر یکدیگر اثر مستقیسم می گذارند.

علاوه براین، مشکل دیگر این مدل فرض سهل انگارانه لزوم بازسازی ضعیف برای نمونه‌های ناهنجار است. در واقع در تمامی مدل‌های پیشین این فرض به طور ضمنی در نظر گرفته شده که اگر مدل با داده های هنجار آموزش ببیند ، لزوما برای داده های ناهنجار نگاشت ضعیفی خواهد داشت در حالی که هیچ قیدی به منظور متمایل کردن مدل به سمت تولید بازسازی ضعیف از نمونه‌های ناهنجار وجود ندارد. در مدل پیشنهادی RCALAD سعی شده است تا تمامی نقاط ضعف اشاره شده در مدل های پیشین پوشش داده بشود.

۴- روش پیشنهادی

در این بخش به طور دقیق تر به بررسی جزئیات هر یک از مشکلات مورد اشاره در فصل قبل پرداخته می‌شود. ابتدا مسئله complete cycle consistency و روش حل آن شرح داده خواهد شد، سپس به بررسی مسئله استلزام بازسازی ضعیف می پردازیم و در انتها مدل پیشنهادی نهایی که با هدف حل هر دو مسئله ذکر شده طراحی شده است معرفی می‌شود.

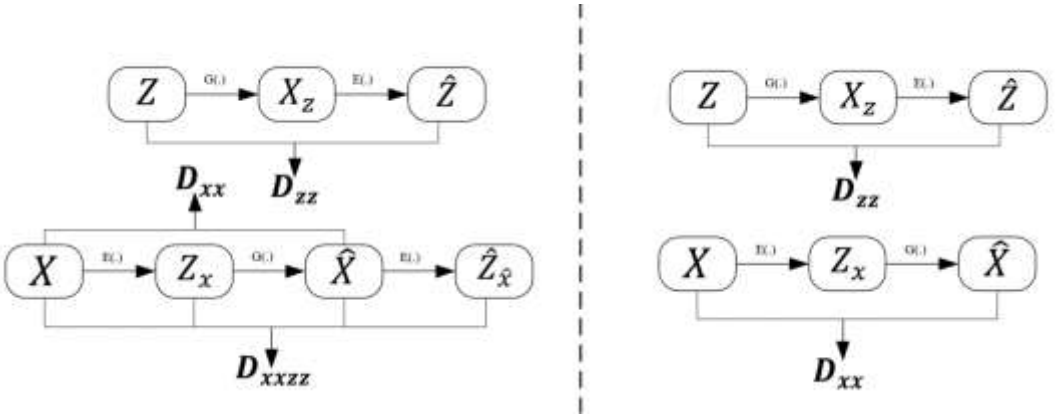
[1] **Lipschitz Constraints**

۴-۱- چرخه پایداری کامل

همانطور که گفته شد، در مدل ALAD پایداری چرخه برای داده ورودی و متغیر فضای نهفته در دو روند مستقل از هم بررسی می‌شود. در واقع روند نزدیک کردن بازسازی متغیرفضای نهان یعنی $\hat{z}$ به خود متغیر z و همچنین نزدیک کردن بازسازی ورودی داده یعنی $\hat{x}$ به خود x به صورت جداگانه صورت می‌پذیرد. لازم به ذکر است در این‌جا مقصود از متغیر z نمونه‌ای از توزیع گوسی است که به عنوان ورودی به شبکه مولد داده می‌شود و ارتباطی با نگاشت ورودی در فضای نهان ندارد .

مسئله Complete Cycle Consistency و یا به اختصار CCC بیان می کند که به ازای هر متغیر x از فضای ورودی اگر ابتدا کدگذار نگاشت معکوس به فضای نهفته را تخمین زند که معادل $E(x) = z_x$ می‌باشد. و در مرحله بعد بازنمایی بدست آمده به شبکه مولد وارد شود تا بازسازی شبکه از متغیر ورودی $G(z_x) = G\big(E(x)\big) = \hat{x}$ تولید شود. سپس همین بازسازی بار دیگر به شبکه کدگذار داده شود تا بازسازی در فضای نهفته نیز محاسبه شود یعنی $E(\hat{x}) = E\big(G(z_x)\big) = \hat{z}_{\hat{x}}$. در این صورت، انتظار منطقی از هر شبکه مبتنی بر بازسازی این است که دو متغیر $\hat{x}$ و x و همچنین دو متغیر $\hat{z}_{\hat{x}}$ و z_x تا حد امکان کمترین اختلاف را داشته باشند. یعنی تعریف مسئله CCC بدین ترتیب می‌شود که، در هر مدل مبتنی بر بازسازی، بایستی به ازای هر داده ورودی و نگاشت آن در فضای نهان، بازسازی ارائه شده توسط شبکه برای هر دو متغیر کمترین خطا و بیشترین شباهت را با آن دو داشته باشد.

در مدل ALAD شباهت میان داده ورودی و بازسازی آن و همچنین شباهت z و بازسازی آن مستقل از هم و در دو چرخه جداگانه بررسی می‌شد و فرض شده بود که مستقل از هم هستند در حالی‌که می‌دانیم این دو چرخه کاملا به یکدیگر وابسته بوده و فرض استقلال این دو مسئله غلط است. در اینجا سعی شده است با بررسی توام متغیرهای موجود در چرخه CCC در تمایزگر جدید D_{xxzz} ، عدم استقلال میان متغیرها مدل شود و جریان اطلاعات موجود در این زنجیره برای بهبود آموزش شبکه و تشخیص هر چه بهتر داده‌های ناهنجار به کار گرفته شود. تفاوت میان ورودی تمایزگر D_{xxzz} و ورودی تمایزگر D_{zz} که در مدل ALAD استفاده شده است در شکل ۱ قابل مشاهده است.



شکل ۱: سمت راست نحوه استفاده از متغیرهای فضای داده ورودی و فضای نهفته در چرخه پایداری شبکه ALAD. سمت چپ نحوه استفاده از اطلاعات یک چرخه کامل در مدل پیشنهادی.

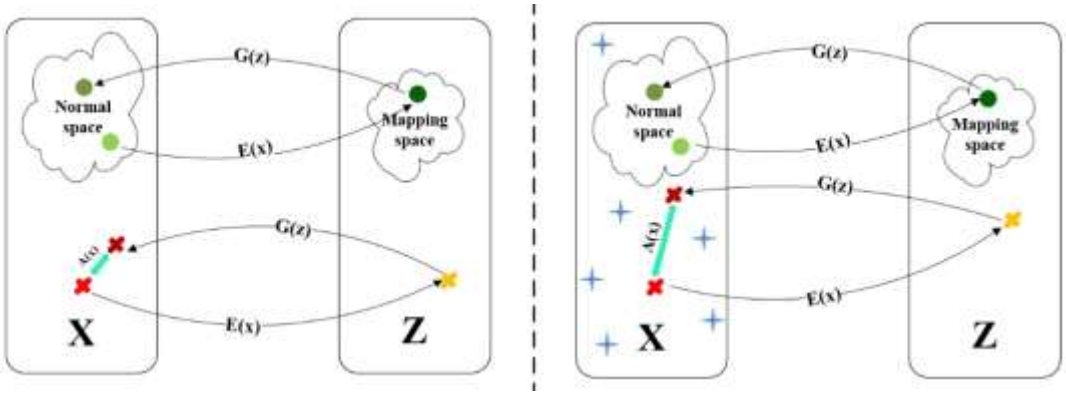
همانطور که در شکل ۱ قابل مشاهده است مدل ALAD از اطلاعات یک چرخه کامل استفاده نمی‌کرد. به منظور استفاده از اطلاعات موجود در یک چرخه کامل متغیر جدید $\hat{z}_{\hat{x}}$ معرفی می‌شود. برای محاسبه این متغیر نگاشت معکوس داده ورودی x به شبکه مولد داده می‌شود و نگاشت معکوس خروجی حاصل مجددا با استفاده از کدگشا محاسبه می‌شود و با این کار یک چرخه کامل از روند دگردیسی داده ورودی حاصل می‌شود.

به منظور تضمین شرط چرخه پایداری کامل از تمایزگر جدید D_{xxzz} با ورودی توام استفاده می‌شود، لازم به ذکر است اثربخشی تمایزگر توامان پیش از این یک مرتبه درALIGAN به اثبات رسیده است. این تمایزگر چهارتایی (x, x, z_x, z_x) به عنوان داده واقعی و از چهارتایی $(x, G(E(x)), z_x, E(G(z_x)))$ به عنوان داده تقلبی ورودی می‌گیرد. در واقع این تمایزگر تلاش می‌کند تا ورودی x و بازسازی ارائه شده از آن توسط شبکه و همینطور نگاشت معکوس تصویر ورودی در فضای نهان یعنی z_x و بازسازی آن توسط کدگذار تا حد امکان به یکدیگر نزدیک باشد تا یک حلقه کامل پایدار توسط مدل ارائه شود و مدل هر چه بهتر آموزش دیده و stabilize شود.

۴-۲- استلزام بازسازی ضعیف

در مدل‌های مبتنی بر بازسازی تاکنون همیشه فرض بر این بوده است که اگر آموزش و بازسازی برای داده‌های هنجار به خوبی انجام بگیرد، بازسازی داده‌های ناهنجار لزوما ضعیف و متفاوت از داده ورودی خواهد بود. اما این پیش‌فرض در بسیاری از موارد صحیح نیست و نمونه بازسازی شده ناهنجار، میزان اختلاف کمی با نمونه ورودی دارد و به همین سبب، تشخیص آن به عنوان نمونه ناهنجار دشوار خواهد بود. در واقع در هیچ یک از مدل‌های پیشین هیچ استلزام یا قید کنترلی برای متمایل کردن مدل به سمت تولید بازسازی ضعیف برای نمونه‌های ناهنجار ارائه نشده است.

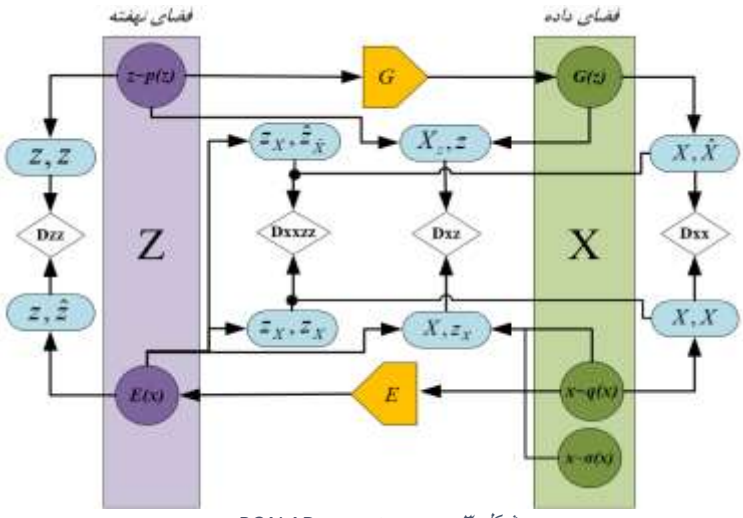
علت وقوع این پدیده نگاشت تنک از فضای داده ورودی به فضای نهفته است. زیرا در فاز آموزش کدگذار تنها نگاشت نمونه‌های هنجار به فضای نهفته را آموزش می‌بیند و در نتیجه فضای متناسب z برای نمونه‌های هنجار به خوبی مدل می‌شود ولی در فاز آزمون با توجه به اینکه مدل تاکنون بقیه فضا از جمله نمونه‌های ناهنجار را ندیده است ممکن است آن را به نقطه‌ای نامعلوم از فضای نهفته نگاشت کند. یعنی در این حالت هیچ اطلاعاتی در خصوص نگاشت برای داده های ناهنجار در دسترس نیست. یک راه‌حل برای این موضوع استلزام نگاشت تمامی فضای ورودی به زیرفضای هنجار نهان می باشد. یعنی در اینجا برای پوشش هرچه بهتر فضای داده ورودی از توزیع نویز با نام $\sigma(x)$ استفاده می‌شود و با نمونه‌گیری از این تابع و متمایل کردن شبکه به سمت تولید بازسازی کلاس داده هنجار شبکه یاد می‌گیرد تا برای طیف به نسبت گسترده‌تری از ورودی‌ها کلاس داده هنجار را بازسازی کند. در این صورت اگر داده ورودی ناهنجار هم باشد مدل آموزش دیده تا بازسازی نزدیک به کلاس داده‌هنجار تولید کند و در نتیجه میان داده ورودی و بازسازی آن فاصله مناسبی ایجاد می‌شود و همین فاصله معیار مناسبی برای تشخیص نمونه‌های ناهنجار خواهد بود. در شکل ۲ با نحوه عملکرد این روال آموزشی آشنا می‌شوید.



شکل ۲: تاثیر حضور توزیع $\sigma(x)$ در روند آموزش مدل. در این شکل x بیانگر فضای داده ورودی و z بیانگر فضای داده ورودی است. نمونه‌ها توسط مولد G از فضای داده ورودی به فضای نهفته نگاشت می‌شوند و وظیفه انجام نگاشت معکوس بر عهده کدگذار E است. دایره‌های سبز رنگ نماد نمونه داده‌های هنجار و ضربدرهای قکد نماد نمونه‌های ناهنجار هستند. بعلاوه آبی رنگ نشانگر نمونه‌های تولید شده توسط توزیع $\sigma(x)$ هستند که در تنها مرحله آموزش مورد استفاده قرار گرفته‌اند. فلش فیروزه‌ای مقدار امتیاز ناهنجاری را نشان می‌دهد . همانطور که در شکل ۳-۵ مشاهده می‌شود در صورت عدم حضور $\sigma(x)$ (در سمت چپ شکل) در روند آموزش، امتیاز ناهنجاری برای نمونه‌های غیرعادی کمتر از زمانی است که از این توزیع استفاده شده است، در تصویر سمت راست، توزیع $\sigma(x)$ مدل را به سمت بازسازی همه نمونه‌ها اعم از ناهنجار و هنجار به سمت توزیع داده‌های هنجار متمایل کرده است.

۴-۳- مدل RCALAD

در این بخش با ترکیب هر دو ایده مطرح شده در بخش های قبل یعنی بکارگیری متغیر جدید $\hat{z}_2$ در تمایزگر D_{xxzz} و همچنین استفاده از توزیع $\sigma(x)$ و افزودن آن‌ها به مدل پایه [4] مدل اصلی پیشنهادی RCALAD معرفی می‌شود. در این شبکه به هر دو مسئله چرخه پایداری کامل و استلزام بازسازی ضعیف به طور همزمان پرداخته شده است و تلاش شده است تا یک چارچوب جامع، کاربردی و سازگار برای تمامی مسائل تشخیص ناهنجاری ارائه شود. شمای کلی مدل پیشنهادی در شکل ۳ قابل مشاهده است.



شکل ۳: ساختار کلی مدل RCALAD.

همانطور که در شکل ۳ مشاهده می‌شود، با هدف کاهش پیچیدگی زمانی، یک کدگذار توام با شبکه مولد در ساختار کلی شبکه عصبی تقابلی آموزش داده می‌شود. نگاشت معکوس از فضای داده ورودی به فضای نهفته به سادگی با تعبیه کدگذار E در معماری پیشنهادی به‌دست می‌آید. در اینجا برای آموزش هم‌زمان هر دو شبکه مولد و کدگذار از یک شبکه تمایزگر توامان با نام D_{xz} استفاده شده است. این تمایزگر بررسی می‌کند که جفت متغیر ورودی متعلق به توزیع داده ورودی x و نقطه متناظر با آن در فضای نهفته $E(x)$ است و یا توسط شبکه مولد $G(z)$ و نمونه‌گیری از فضای نهفته z تولید شده است. به منظور ارضای شرط پایداری حلقه در فضای داده ورودی از تمایزگر D_{xx} و D_{zz} استفاده شده است تا هر نمونه و بازسازی متناظر با آن به طور مستقل بهبود یافته وشبیه شوند. تمایزگر D_{xxzz} با هدف استفاده از تمامی اطلاعات موجود در یک چرخه کامل به صورت توام اضافه شده است. یعنی در کنار بررسی هر دو متغیر و بازسازی آن‌ها در همان فضا، توزیع توامان چهارتایی آن‌ها در روند تشخیص نمونه ناهنجار به کار گرفته شود تا شبکه به وضعیت داده ورودی در حین نگاشت‌های متوالی دسترسی داشته باشد و اطلاعات بیشتری برای تمییز داده ها در دسترس داشته باشد. این شبکه وظیفه تمایز بین نمونه‌های چهارتایی (x, x, z_x, z_x) و $(x, G(E(x)), z_x, E(G(z_x)))$ را دارد و تلاش می‌کند تا x و بازسازی ارائه شده توسط شبکه و همینطور نگاشت تصویر ورودی در فضای نهان z_x و بازسازی خروجی شبکه مولد توسط کدگذار $E(G(z_x))$ تا حد امکان به طور وابسته و توامان به یکدیگر نزدیک کند. بلوک $\sigma(x)$ به منظور پوشش حداکثری فضای نهفته به این مدل اضافه شده است. هدف از تعبیه این بلوک تولید نمونه‌های جدید در فضای داده ورودی و سپس نگاشت آن به فضای نهفته متناسب با داده هنجار است. در نهایت تابع هدف مدل پیشنهادی به صورت زیر است.

$$\begin{aligned} min_{G,E} max_{D_{xxzz},D_{xz},D_{xx},D_{zz}} V_{RCALAD}(D_{xxzz}, D_{xz}, D_{xx}, D_{zz}, E, G) \\ = V_{ALAD} + \mathbb{E}_{x\sim\sigma(x)}\left[\log\left(1 - D_{xz}(x, E(x))\right)\right] + \mathbb{E}_{x\sim q(x)}\left[\log D_{xxzz}(x, x, E(x), E(x))\right] \\ + \mathbb{E}_{x\sim q(x)}\left[1 - log\, D_{xxzz}\left(x, G(E(x)), E(x), E\left(G(E(x))\right)\right)\right] \end{aligned} \quad (۷)$$

۴-۴- تشخیص ناهنجاری

هدف اصلی از ارائه مدل پیشنهادی در این مقاله، تشخیص ناهنجاری بر اساس بازسازی داده ورودی است. در این مدل هدف نهایی بازسازی دقیق و شبیه برای داده‌های هنجار و بازسازی ضعیف و متفاوت برای نمونه ناهنجار است. یکی از عناصر کلیدی در تشخیص ناهنجاری، تعریف امتیاز ناهنجاری با هدف محاسبه فاصله میان نمونه ورودی و بازسازی ارائه شده توسط شبکه است.[4].

اولین امتیاز ناهنجاری ارائه شده در این مقاله، در این کار $A_{fm}(x)$ نام دارد. در این امتیاز برای محاسبه فاصله میان نمونه‌ها و بازسازی آن‌ها، از فضای ویژگی موجود در تمایزگر D_{xxzz} استفاده می‌شود. به این منظور خروجی logit های لایه یکی مانده به آخر، به عنوان ویژگی استفاده می‌شوند. امتیاز ناهنجاری مورد استفاده به صورت زیر و با استفاده از خطای بازسازی نرم یک و مطابق معادله زیر تعریف می‌شود.

$$A_{fm}(x) = \|f_{xxzz}(x, x, E(x), E(x)) - f_{xxzz}(x, G(E(x)), E(x), E(G(E(x))))\|_1 \quad (۸)$$

در این معادله $f(\cdot)$ بیانگرتابع فعالیت لایه یکی مانده به آخر در ساختار تمایزگر D_{xxzz} است. مفهوم بکارگرفته شده پشت تعریف این امتیاز، بکارگیری از میزان اطمینان تمایزگر از کیفیت بازسازی‌های ارائه شده توسط شبکه است که اگر خوب انجام شده باشد در واقع نمونه متعلق به

داده آموزش دیده شبکه یا همان توزیع داده هنجار است. بدین ترتیب هر چه مقدار این معیار بیشتر باشد، اختلاف بازسازی‌ها بیشتر بوده و احتمال ناهنجاری بودن آن داده ورودی بیشتر است .

دومین امتیاز ارائه‌شده در این مقاله با هدف بیشینه کردن بهره‌گیری از اطلاعات موجود در مدل برای تشخیص ناهنجاری می‌باشد. در این بخش معیار A_{all} تعریف شده است. این امتیاز از جمع خروجی هر سه تمایزگر D_{xx} ، D_{zz} و D_{xxzz} تشکیل شده است. در واقع چون در اینجا تمام تمایزگرهای موجود در مدل پیشنهادی تنها روی نمونه‌های هنجار آموزش دیده‌اند و بازسازی برای تمامی فضای داده ورودی به سمت فضای داده هنجار متمایل شده است، پس انتظار می‌رود تصویر بازسازی شده نمونه ناهنجار و همچنین بازنمایی آن در فضای نهفته که توسط کدگذار تولید می‌شود، بسیار متفاوت از ورودی باشد و تمایزگرهای موجود در مدل به راحتی این ورودی‌های ناهنجار را شناسایی کنند. بیان ریاضی این معیار در معادله زیر آورده شده است.

$$A_{all}(x) = D_{xxzz}(x, \hat{x}, z_x, \hat{z}_{\hat{x}}) + D_{xx}(x, \hat{x}) + D_{zz}(z_x, \hat{z}_{\hat{x}}) \quad (۹)$$

حال مسئله قابل بررسی این موضوع می‌باشد که آیا معیار A_{all} حاوی اطلاعات کافی برای تشخیص داده‌های هنجار از ناهنجار می‌باشد یا خیر. پاسخ به این سوال در حالت کلی بله می‌باشد زیرا این تمایزگرها در طی فرایند آموزش یاد می‌گیرند که به اختلاف میان دوتایی (x, x) و $(x, \hat{x})$ و همچنین دوتایی (z_x, z_x) و $(z_x, \hat{z}_{\hat{x}})$ توجه کنند یعنی هر چه $\hat{x}$ از x و یا $\hat{z}_{\hat{x}}$ از z_x فاصله بگیرد، تشخیص آن برای تمایزگرها ساده‌تر می‌شود. در مدل پیشنهادی با افزودن توزیع $\sigma(x)$ و متمایل سازی تمامی بازسازی به سمت توزیع داده‌های هنجار، اختلاف بازسازی برای داده‌های ناهنجار زیادتر هم شده و عملاً خروجی تمایزگرها یک معیار قابل اتکا برای تشخیص ناهنجاری می‌شود. در نهایت الگوریتم محاسبه امتیازهای ناهنجاری پیشنهادی مطابق الگوریتم زیر قابل مشاهده است.

Algorithm 2 Regularized Complete Adverarially Learned Anomaly Detection

Input $x \sim p_{X_{\text{Test}}}(x), E, G, D_{xx}, D_{zz}, D_{xxzz}, f_{xxzz}$ where f_{xxzz} is the feature layer of , D_{xxzz}
Output $A_{all}(x)$, $A_{fm}(x)$, **where A is the anomaly score**

procedure INFERENCE
- $z_x \leftarrow E(x)$ — Encode samples, Construct latent Embedding
- $\hat{x} \leftarrow G(z_x)$ — Reconstruct samples
- $\hat{z}_{\hat{x}} \leftarrow E(\hat{x})$ — Reconstruct latent Embedding
- $A_{fm}(x) \leftarrow \|f_{xxzz}(x, x, z_x, z_x) - f_{xx}(x, \hat{x}, z_x, \hat{z}_{\hat{x}})\|_1$
- $A_{all}(x) \leftarrow D_{xxzz}(x, \hat{x}, Z_x, \hat{z}_{\hat{x}}) + D_{xx}(x, \hat{x}) + D_{zz}(Z_x, \hat{z}_{\hat{x}})$
- return $A_{all}(x)$, $A_{fm}(x)$

end procedure

الگوریتم ۱: روند محاسبه‌ی امتیاز ناهنجاری.

۵- آزمایش‌ها

در این بخش به ارزیابی عملکرد مدل پیشنهادی در مقایسه با مدل‌های مطرح حوزه تشخیص ناهنجاری که در بخش ۲ به تفصیل بررسی شد، می‌پردازیم. با هدف منصفانه بودن آزمایش‌ها، نتایج گزارش شده برای تمامی مدل‌ها بر روی دادگان جدولی حاصل از میانگین ده اجرا و برای هر کلاس از دادگان تصویری میانگین سه اجرا می‌باشد.

۵-۱- دادگان

برای سنجش عملکرد مدل پیشنهادی و بررسی کارایی آن از جنبه‌های مختلف از مجموعه داده‌هایی با ویژگی متفاوت استفاده می‌شود. روش پیشنهادی روی مجموعه داده‌های تصویری و جدولی موجود آزمایش شده است. برای دادگان جدولی از چهار مجموعه داده kddcup99 ، arrhythmia، thyroid و musk استفاده شده است. دادگان kddcup99 دادگان مرتبط با نفوذ در شبکه است arrhythmia .که یک مجموعه پزشکی مرتبط با آریتمی قلبی با ۱۶ کلاس است thyroid .یک دادگان سه کلاسی مرتبط با بیماری تیروییید است. دادگان musk با هدف دسته‌بندی شش کلاسی روی مشک مولکولی ایجاد شده است. در مجموعه داده‌های معرفی شده به ترتیب ۲۰، ۱۵، ۲.۵ و ۳.۲ درصد از داده‌ها نمونه‌های ناهنجار هستند بنابراین در مرحله آزمون پس از محاسبه امتیاز ناهنجاری همین نسبت از دادگان که بیشترین مقدار امتیاز ناهنجاری را دارا می‌باشند به عنوان ناهنجاری دسته‌بندی می‌شوند. ارزیابی مدل پیشنهادی روی این مجموعه داده‌ها با محاسبه معیارهای F1 ، Recallو Precisionصورت می‌پذیرد. برای مجموعه داده تصویری دو دادگان CIFAR10 و SVHN در نظر گرفته شده‌اند. هر دو این دادگان ده کلاسی

هستند و هربار یک کلاس به عنوان کلاس هنجار و سایر ۹ کلاس به عنوان کلاس ناهنجار در نظر گرفته می‌شود. معیار مورد استفاده برای ارزیابی مدل روی دادگان تصویری area under the receiver operating curve به اختصار AUROC است.

برای تمامی دادگان مورد استفاده ۸۰ درصد دادگان به عنوان داده آموزشی و ۲۰ درصد به عنوان داده آزمون انتخاب می‌شود. ۲۵ درصد از داده آموزشی به عنوان داده ارزیابی (validation) انتخاب می‌شود. لازم به ذکر است در مرحله آموزش همه نمونه‌های ناهنجار از داده آموزشی حذف می‌شود.

۵-۲- آزمایش‌ها روی دادگان جدولی

نتایج ارزیابی مدل پیشنهادی RCALAD روی دادگان جدولی kddcup99، arrhythmia، thyroid و musk در جدول ۱ خلاصه شده است. به جز مدل ALAD که بر روی هر مجموعه داده اجرا و نتایجش گزارش شده، برای سایر مدل‌ها از نتایج موجود [26] استفاده شده است. ساختارهای مورد استفاده در شبکه مولد، تمایزگر و کدگذار همگی لایه‌های کاملا متصل استاندارد با توابع فعالساز غیرخطی هستند. لازم به ذکر است در این مرحله از توزیع $N(0,I)$ به عنوان $\sigma(x)$ استفاده می‌شود.

	KDDCUP			Arrhythmia			Thyroid			Musk		
Model	Prec.	Recall	F_1	Prec.	Recall	F_1	Prec.	Recall	F_1	Prec.	Recall	F_1
IF	92.16	93.73	92.94	51.47	54.69	53.03	**70.13**	**71.43**	**70.27**	47.96	47.72	47.51
OC-SVM	74.57	85.23	79.54	53.97	40.82	45.18	36.39	42.39	38.87	–	–	–
DSEBMr	85.12	64.72	73.28	15.15	15.13	15.10	4.04	4.03	4.03	–	–	–
DSEBMe	86.19	64.46	73.99	46.67	45.65	46.01	13.19	13.19	13.19	–	–	–
AnoGAN	87.86	82.97	88.65	41.18	43.75	42.42	44.12	46.87	45.45	3.06	3.10	3.10
DAGMM	92.97	94.22	93.69	49.09	50.78	49.83	47.66	48.34	47.82	–	–	–
ALAD	94.27	**95.77**	95.01	50.00	53.13	51.52	22.92	21.57	22.22	58.16	59.03	58.37
DSVDD	89.81	94.97	92.13	35.32	34.35	34.79	22.22	23.61	23.29	–	–	–
RCALAD	**95.36**	95.62	**95.49**	**58.82**	**62.50**	**60.60**	53.76	51.53	52.62	**62.96**	**63.33**	**63.14**
error bar	0.28	0.29	0.28	6.6	6.8	5.8	4.3	2.7	2.8	5.06	2.53	2.62

جدول ۱: نتایج خروجی مدل پیشنهادی در مقایسه با مدل‌های پایه بر روی مجموعه داده‌های جدولی.

برای مقایسه شفاف میان مدل‌های مختلف از error bar در ردیف آخر جدول ۱ استفاده شده است. همانطور که در این جدول قابل مشاهده است مدل پیشنهادی روی دادگان arrhythmia و musk نسبت به سایر مدل‌ها بسیار موفق عمل کرده است. روی دادگان KDD بر اساس معیار F1 بهترین مدل است ولی روی دادگان thyroid با توجه به عملکرد کم‌نظیر مدل IF در رتبه دوم قرار می‌گیرد. علت این پدیده می‌تولند در جنس داده های این دیتاست باشد؛ زیرا در این مجموعه داده تعداد زیادی ویژگی وجود دارد که تنها تعداد کمی از آن ها informative هستند و لذا نتایج مدل‌های کلاسیک مانند IF که مبتنی بر انتخاب ویژگی هستند، بهتر است. یک ایده برای بهبود نتایج مدل پیشنهادی، بکارگیری مدل‌هایی نظیر IFدر مرحله پیش پردازش برای انتخاب ویژگی های موثرتر برای آموزش مدل است.

۵-۳- آزمایش‌ها روی دادگان تصویری

در این قسمت عملکرد مدل پیشنهادی روی دادگان تصویری در دو جدول مستقل از هم بررسی می‌شود. همانطور که در جدول ۲ و ۳ قابل مشاهده است، مدل پیشنهادی روی دادگان CIFAR10 بهبود قابل توجهی ایجاد کرده است.

جدول ۲: نتایج خروجی مدل پیشنهادی در مقایسه با مدل های پایه برروی مجموعه داده cifar10.

Normal	DCAE	DSEBM	DAGMM	IF	AnoGAN	ALAD	RCALAD
Airplane	59.1 ± 5.1	41.4 ± 2.3	56.0 ± 6.9	60.1 ± 0.7	67.1 ± 2.5	64.7 ± 2.6	$\mathbf{72.8 \pm 0.8}$
auto.	$\mathbf{57.4 \pm 2.9}$	57.1 ± 2.0	56.0 ± 6.9	50.8 ± 0.6	54.7 ± 3.4	45.7 ± 0.8	50.2 ± 0.3
Bird	48.9 ± 2.4	61.9 ± 0.1	53.8 ± 4.0	49.2 ± 0.4	52.9 ± 3.0	67.0 ± 0.7	$\mathbf{72.6 \pm 0.2}$
Cat	58.4 ± 1.2	50.1 ± 0.4	51.2 ± 0.8	55.1 ± 0.4	54.5 ± 1.9	59.2 ± 1.1	$\mathbf{64.2 \pm 0.9}$
Deer	54.0 ± 1.3	73.2 ± 0.2	52.2 ± 7.3	49.8 ± 0.4	65.1 ± 3.2	72.7 ± 0.6	$\mathbf{74.9 \pm 0.5}$
Dog	$\mathbf{62.2 \pm 1.8}$	60.5 ± 0.3	49.3 ± 3.6	58.5 ± 0.4	60.3 ± 2.6	52.8 ± 1.2	60.1 ± 1.1
Frog	51.2 ± 5.2	68.4 ± 0.3	64.9 ± 1.7	42.9 ± 0.6	58.5 ± 1.4	69.5 ± 1.1	$\mathbf{75.3 \pm 0.4}$
Horse	58.6 ± 2.9	53.3 ± 0.7	55.3 ± 0.8	55.1 ± 0.7	$\mathbf{62.5 \pm 0.8}$	44.8 ± 0.4	56.6 ± 0.2
Ship	$\mathbf{76.8 \pm 1.4}$	73.9 ± 0.3	51.9 ± 2.4	74.2 ± 0.6	75.8 ± 4.1	73.4 ± 0.4	$\mathbf{77.5 \pm 0.3}$
Truck	$\mathbf{67.3 \pm 3.0}$	63.6 ± 3.1	54.2 ± 5.8	58.9 ± 0.7	66.5 ± 2.8	43.2 ± 1.3	52.6 ± 0.6
Mean	59.4	60.3	54.4	55.5	61.8	59.3	**65.7**

روی دادگان SVHN نیز (با اختلاف کمتری نسبت به مدل پایهALAD) بهترین عملکرد را داشته است.

جدول ۳: نتایج خروجی مدل پیشنهادی در مقایسه با مدل های پایه برروی مجموعه داده SVHN.

Normal	OCSVM	DSEBMr	DSEBMe	IF	ANOGAN	ALAD	RCALAD
0	52.0 ± 1.6	56.1 ± 0.2	53.4 ± 1.8	53.0 ± 0.6	57.3 ± 0.4	58.7 ± 0.9	$\mathbf{60.4 \pm 0.1}$
1	48.6 ± 5.3	52.3 ± 0.9	52.1 ± 0.3	51.2 ± 0.9	57.0 ± 0.8	$\mathbf{62.8 \pm 1.7}$	59.2 ± 0.3
2	49.7 ± 7.7	51.9 ± 0.8	51.8 ± 0.4	52.3 ± 0.1	53.1 ± 0.4	$\mathbf{55.2 \pm 2.3}$	54.9 ± 0.1
3	50.9 ± 1.4	51.8 ± 0.4	51.7 ± 0.5	52.2 ± 0.3	52.6 ± 0.4	53.8 ± 3.3	$\mathbf{55.8 \pm 1.9}$
4	48.4 ± 5.2	52.5 ± 0.1	52.4 ± 0.2	49.1 ± 0.6	53.9 ± 0.5	58.0 ± 0.1	$\mathbf{58.5 \pm 0.2}$
5	51.1 ± 2.6	52.4 ± 2.3	52.3 ± 2.6	52.4 ± 0.8	52.8 ± 0.1	56.1 ± 0.9	$\mathbf{56.2 \pm 0.4}$
6	50.1 ± 3.9	52.1 ± 1.8	52.2 ± 1.8	51.8 ± 0.2	53.2 ± 0.0	57.4 ± 0.6	$\mathbf{59.4 \pm 0.5}$
7	49.6 ± 1.3	53.4 ± 0.9	55.3 ± 1.1	52.0 ± 0.4	55.0 ± 0.0	$\mathbf{58.8 \pm 0.3}$	58.0 ± 0.4
8	45.0 ± 4.2	51.9 ± 0.3	52.5 ± 0.6	52.3 ± 0.8	52.2 ± 0.7	$\mathbf{55.2 \pm 0.4}$	$\mathbf{56.1 \pm 0.5}$
9	52.5 ± 3.9	55.8 ± 1.7	52.7 ± 1.4	53.7 ± 0.6	53.1 ± 0.1	57.3 ± 0.6	$\mathbf{58.3 \pm 0.2}$
Mean	50.2	52.9	52.4	51.6	54.0	57.3	**57.7**

۵-۴- مطالعات فرسایشی

در این قسمت کارکرد هر یک از اجزای اضافه شده به مدل پایه را روی دادگان CIFAR10 و SVHN بررسی می‌کنیم. آزمایش‌ها در این قسمت در زمان حضور و عدم حضور تمایزگر D_{xxzz} و همچنین حضور و عدم حضور $\sigma(x)$ تکرار می‌شود.

جدول ۴: تاثیر بخش‌های مختلف پیشنهادی در بهبود نتایج دادگان جدولی.

Model	AUROC
SVHN	
Baseline (ALAD)	0.573 ± 0.016
Baseline + D_{xxzz} (CALAD)	0.576 ± 0.014
Baseline + $\sigma(x)$ (RALAD)	0.568 ± 0.018
Baseline + D_{xxzz} + $\sigma(x)$ (RCALAD)	$\mathbf{0.577 \pm 0.019}$
CIFAR-10	
Baseline (ALAD)	0.593 ± 0.017
Baseline + D_{xxzz} (CALAD)	0.634 ± 0.018
Baseline + $\sigma(x)$ (RALAD)	0.642 ± 0.012
Baseline + D_{xxzz} + $\sigma(x)$ (RCALAD)	$\mathbf{0.657 \pm 0.016}$

همانطور که در جدول۴ مشخص است، مدل پیشنهادی RCALAD به بالاترین کارایی در حضور هر دو بخش دست می یابد. در بررسی نقش تمایزگر D_{xxzz} می توان گفت این تمایزگر روی دادگان CIFAR10 دقت را تا اندازه خوبی بهبود داده اســـت ولی بهبود قابل توجهی روی دادگان SVHN ایجاد نکرده است. در خصوص نقش توزیع $\sigma(x)$ می توان گفت که این توزیع روی دادگان CIFAR10 عملکرد مناسبی داشته و معیار AUROC را بهبود داده است. اما در دادگان SVHN معیار AUROC را به میزان کمی به نسبت مدل پایه کاهش داده است اما وجودش در مدل نهایی سبب استخراج اطلاعات جدید و نگاه جامع تر می شود.

۵-۵- ارزیابی کفایت تمایزگر D_{xxzz}

با اضافه کردن تمایزگر D_{xxzz} آیا نیازی به تمایزگر D_{xx} و D_{zz} است یا خیر؟ برای پاسخگویی مناسب به این سوال باید جدول زیر را مشاهده کرد. در واقع در این بخش علاوه بر سوال فوق به بررسی نتیجه افزودن تمایزگر D_{xxzz} در مدل‌های پایه نظیر ALI و ALICE نیز پرداخته شده است.

جدول ۵: ارزیابی عملکرد مدل در حضور/عدم حضور هر یک از اجزا.

Model	D_{zz}	D_{xx}	D_{xxzz}	Prec.	Recall	F_1
KDD99						
ALAD	√	√	×	0.942 ± 0.008	$\mathbf{0.957 \pm 0.006}$	0.950 ± 0.007
ALI + D_{xxzz}	×	×	√	0.938 ± 0.007	0.951 ± 0.010	0.944 ± 0.009
ALI + D_{zz} + D_{xxzz}	√	×	√	0.946 ± 0.005	0.955 ± 0.004	0.950 ± 0.004
ALICE + D_{xxzz}	×	√	√	0.941 ± 0.005	0.954 ± 0.008	0.947 ± 0.006
CALAD	√	√	√	$\mathbf{0.959 \pm 0.004}$	0.957 ± 0.007	$\mathbf{0.958 \pm 0.005}$
RCALAD	√	√	√	0.953 ± 0.007	0.956 ± 0.005	0.954 ± 0.006
Arrhythmia						
ALAD	√	√	×	0.500 ± 0.049	0.531 ± 0.047	0.515 ± 0.048
ALI + D_{xxzz}	×	×	√	0.522 ± 0.054	0.529 ± 0.049	0.525 ± 0.052
ALI + D_{zz} + D_{xxzz}	√	×	√	0.571 ± 0.033	0.582 ± 0.028	0.576 ± 0.031
ALICE + D_{xxzz}	×	√	√	0.543 ± 0.052	0.561 ± 0.044	0.551 ± 0.048

CALAD	√	√	√	0.574 ± 0.021	0.605 ± 0.022	0.575 ± 0.021
RCALAD	√	√	√	$\mathbf{0.588 \pm 0.42}$	$\mathbf{0.625 \pm 0.41}$	$\mathbf{0.606 \pm 0.41}$

همانطور که در جدول ۵ مشـــاهده میشـــود، مطابق انتظار از نتایج تئوری، افزودن تمایزگر D_{xxzz} به چارچوب کلی و در کنار دیگر تمایزگرها بالاترین کارایی را داشــته اســت. پس از آن حذف D_{xx} ضــربه کمتری به مدل می‌زند زیرا بخشــی از اطلاعاتی که اســتخراج می‌کند، توســط تمایزگر D_{xxzz} پوشــش داده می‌شــود. اما با توجه به اینکه D_{zz} در یک چرخه مســتقل میزان شــباهت z و بازســازی آن را مورد بررســی قرار می‌دهد طبیعی اســت که حذف آن میزان دقت را کاهش دهد. همان‌طور که دیده می‌شــود، نتیجه این بخش این اســت که این ســه تمایزگر در کنار هم بیشترین کارایی را دارند و تمایزگر D_{xxzz} به تنهایی تمامی جنبه‌ها را پوشش نمی‌دهد.

۵-۶- تشخیص ناهنجاری

در این قسمت امتیازهای ناهنجاری معرفی شده در این مقاله مورد ارزیابی قرار گرفته و با امتیازهای ناهنجاری معرفی شده در کارهای پیشین [4] مقایســه می‌شــود. در این‌جا خروجی خام تمایزگرها با عنوان لاجیت نامگذاری می‌شــود، همچنین خروجی لایه پنهان قبل از لایه لاجیت، ویژگی نامیده می‌شود. در اینجا برای محاســبه امتیاز ناهنجاری از متغیرهای موجود در ســاختار تمایزگر D_{xxzz} اســتفاده می‌شــود. بیان ریاضــی تمامی امتیازهای ناهنجاری موجود به شرح زیر است.

$$
\begin{aligned}
&A_{L_1}(x) = \|x - \hat{x}\|_1 \\
&A_{L_2}(x) = \|x - \hat{x}\|_2 \\
&A_{Logits}(x) = \log\left(D_{xx}(x, \hat{x})\right) \\
&A_{Features}(x) = \|f_{xx}(x, x) - f_{xx}(x, \hat{x})\|_1 \\
&A_{fm}(x) = \|f_{xxzz}(x, x, z_x, z_x) - f_{xxzz}(x, \hat{x}, z_x, \hat{z}_{\hat{x}})\|_1 \\
&A_{all}(x) = D_{xxzz}(x, \hat{x}\ , z_x\ , \hat{z}_{\hat{x}}) + D_{zz}(z_x, \hat{z}_{\hat{x}}) + D_{xx}(x, \hat{x}\ )
\end{aligned}
\qquad (۱۰)
$$

جدول ۶: مقایسه عملکرد امتیازهای ناهنجاری پیشنهادی با سایر امتیازها روی دادگان تصویری.

Model	Precision	Recall	F1 score
KDD99			
A_{L_1}	0.9081 ± 0.0638	0.9108 ± 0.0638	0.9094 ± 0.0638
A_{L_2}	0.9011 ± 0.0155	0.9004 ± 0.0157	0.9007 ± 0.0156
A_{Logits}	0.9169 ± 0.0162	0.9168 ± 0.0164	0.9168 ± 0.0163
$A_{Features}$	0.9127 ± 0.0029	0.9177 ± 0.0039	0.9151 ± 0.0034
A_{fm}	$\mathbf{0.9327 \pm 0.0017}$	$\mathbf{0.9377 \pm 0.0017}$	$\mathbf{0.9301 \pm 0.0017}$
A_{all}	0.9231 ± 0.0018	0.9207 ± 0.0018	0.9218 ± 0.0018
Arrhythmia			
A_{L_1}	0.3529 ± 0.0148	0.3750 ± 0.0164	0.3636 ± 0.0256
A_{L_2}	0.3529 ± 0.0107	0.3750 ± 0.0108	0.3636 ± 0.0107
A_{Logits}	0.5588 ± 0.0334	0.5937 ± 0.0386	0.5757 ± 0.0359
$A_{Features}$	0.2325 ± 0.0029	0.2500 ± 0.0029	0.2424 ± 0.0029
A_{fm}	0.4411 ± 0.0013	0.4687 ± 0.0013	0.4545 ± 0.0013
A_{all}	$\mathbf{0.6176 \pm 0.0208}$	$\mathbf{0.6562 \pm 0.0221}$	$\mathbf{0.6363 \pm 0.0214}$
Thyroid			
A_{L_1}	0.4981 ± 0.0028	0.4908 ± 0.0024	0.4994 ± 0.0024
A_{L_2}	0.5011 ± 0.0330	0.5004 ± 0.0318	0.5007 ± 0.0324
A_{Logits}	0.4969 ± 0.0142	0.4968 ± 0.0144	0.4968 ± 0.0143
$A_{Features}$	0.5127 ± 0.0119	0.5177 ± 0.0119	0.5151 ± 0.0119
A_{fm}	0.5227 ± 0.0083	0.5123 ± 0.0083	0.5174 ± 0.0083
A_{all}	$\mathbf{53.76 \pm 0.0029}$	$\mathbf{51.53 \pm 0.0029}$	$\mathbf{52.62 \pm 0.0029}$
Musk			
A_{L_1}	0.5979 ± 0.0103	0.5931 ± 0.0109	0.5954 ± 0.0106
A_{L_2}	0.6008 ± 0.0021	0.6018 ± 0.0028	0.6013 ± 0.0024
A_{Logits}	0.5868 ± 0.0124	0.5897 ± 0.0127	0.5882 ± 0.0125
$A_{Features}$	0.5824 ± 0.0011	0.5883 ± 0.0019	0.5883 ± 0.0015
A_{fm}	0.6111 ± 0.0481	0.6187 ± 0.0468	0.6148 ± 0.0474

A_{all}	$\mathbf{62.96 \pm 0.0013}$	$\mathbf{63.33 \pm 0.0013}$	$\mathbf{63.14 \pm 0.0013}$

همانطور که در جدول ۶ مشـاهده می‌شـود، روی دادگان جدولی، خروجی خام تمایزگر D_{xxzz} یعنی A_{all} دارای بهترین نتایج به نسـبت سـایر امتیازهای ناهنجاری است. مقدار امتیازها روی دادگان تصویری، مطابق جدول ۷ به شرح زیر است.

جدول ۷: مقایسه عملکرد امتیازهای ناهنجاری پیشنهادی با سایر امتیازها روی دادگان تصویری.

Anomaly Score	AUROC
SVHN	
A_{L_1}	0.5778 ± 0.0141
A_{L_2}	0.5636 ± 0.0251
A_{Logits}	0.5369 ± 0.0785
$A_{Features}$	0.5763 ± 0.0367
A_{fm}	0.5768 ± 0.0251
A_{all}	0.5778 ± 0.0161
CIFAR-10	
A_{L_1}	63.41 ± 0.0321
A_{L_2}	63.27 ± 0.0782
A_{Logits}	62.97 ± 0.0643
$A_{Features}$	63.12 ± 0.0368
A_{fm}	64.77 ± 0.0227
A_{all}	65.73 ± 0.0194

همانطور که در جدول ۷ مشخص است عملکرد امتیاز مبتنی بر ویژگی‌ها یعنی A_{fm} روی دادگان تصویری بسیار مناسب است، این تفاوت در عملکرد امتیازها می‌تواند به دلیل تفاوت در تعداد ویژگی‌های این دو جنس مجموعه داده باشـد، در واقع با توجه به اینکه تعداد ویژگی‌ها روی دادگان جدولی به نسبت دادگان تصویری کمتر است، تمایزگر D_{xxzz} قادر به استخراج و تشخیص مناسب نمونه‌های ناهنجار است، اما در مجموعه داده تصویری، خروجی لایه پنهان قبل از لایه لاجیت حاوی اطلاعات غنی تر برای distinguish بین داده های هنجار و ناهنجار است که سبب بهبود عملکرد امتیاز A_{fm} شده است.

۶- جمع‌بندی

در این کار یک مدل جدید مبتنی بر شبکه‌های عصبی تقابلی به منظور تشخیص ناهنجاری معرفی شد. در مدل پیشنهادی از یک کدگذار برای نگاشت معکوس از فضای داده ورودی استفاده می‌شود و برای ارضای شرط چرخه پایداری از تمایزگر D_{xx} کمک گرفته می‌شود. به منظور پایدارسازی روند آموزش شبکه مولد تقابلی تمایزگر D_{zz} در ساختار تقابلی مدل بهره گرفته می‌شود. با هدف استفاده از اطلاعات یک چرخه کامل در مدل پیشنهادی متغیر $\hat{z}_{\hat{x}}$ معرفی و در نتیجه تمایزگر D_{xxzz} در مدل پیشنهادی جای داده شد. علاوه بر این، برای متمایل‌سازی خروجی شبکه به سمت توزیع داده نرمال، از توزیع $\sigma(x)$ استفاده می‌شود. نتایج حاصل از آزمایش‌ها بیانگر اثربخشی مدل پیشنهادی در زمینه تشخیص ناهنجاری و همچنین برتری آن نسبت به سایر مدل‌های state of the art روی دادگان جدولی و تصویری است. علی رغم نتایج چشمگیر و درخشان مدل پیشنهادیRCALAD ، این مدل همانند دیگر مدل های مبتنی بر GAN، از مشکل robust نبودن نتایج نسبت به کلاس ناهنجاری رنج میبرد و در ادامه کار، می توان با اعمال روش های موجود در این حوزه[27]، مدل را تقویت کرد.

مراجع

[1] X. Shu, L. Cheng, and S. J. Stolfo, "Anomaly Detection as a Service."

[2] C. Jiang, J. Song, G. Liu, L. Zheng, and W. Luan, "Credit Card Fraud Detection: A Novel Approach Using Aggregation Strategy and Feedback Mechanism," *IEEE Internet Things J.*, vol. 5, no. 5, pp. 3637–3647, 2018, doi: 10.1109/JIOT.2018.2816007.

[3] X. Dai and M. Bikdash, "Distance-based outliers method for detecting disease outbreaks using social media," *Conf. Proc. - IEEE SOUTHEASTCON*, vol. 2016-July, 2016, doi: 10.1109/SECON.2016.7506752.

[4] H. Zenati, M. Romain, C. S. Foo, B. Lecouat, and V. Chandrasekhar, "Adversarially Learned Anomaly Detection," *Proc. - IEEE Int. Conf. Data Mining, ICDM*, vol. 2018-Novem, pp. 727–736, 2018, doi: 10.1109/ICDM.2018.00088.

[5] I. Goodfellow *et al.*, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems*, Oct. 2014, pp. 2672–2680, doi: 10.1109/ICCVW.2019.00369.

[6] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional

generative adversarial networks," *4th Int. Conf. Learn. Represent. ICLR 2016 - Conf. Track Proc.*, pp. 1–16, 2016.

[7] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, "Generative Adversarial Networks: An Overview," *IEEE Signal Process. Mag.*, vol. 35, no. 1, pp. 53–65, 2018, doi: 10.1109/MSP.2017.2765202.

[8] T. Schlegl, P. Seeb, S. Waldstein, U. Schmidt-Erfurth, and G. Langs, "Unsupervised Anomaly Detection with Generative Adversarial Networks to Guide Marker Discovery," *Int. Confrence Inf. Process. Med. Imaging*, vol. 2, pp. 146–157, 2017, doi: 10.1007/978-3-319-59050-9.

[9] R. Kaur and S. Singh, "A survey of data mining and social network analysis based anomaly detection techniques," *Egypt. Informatics J.*, vol. 17, no. 2, pp. 199–216, 2016, doi: 10.1016/j.eij.2015.11.004.

[10] A. Zimek, E. Schubert, and H. Kriegel, "REVIEW A Survey on Unsupervised Outlier Detection in High-Dimensional Numerical Data," *Signal Processing*, vol. 99, pp. 215–249, 2012, doi: 10.1002/sam.

[11] M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Processing*, vol. 99, pp. 215–249, 2014, doi: 10.1016/j.sigpro.2013.12.026.

[12] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Piatt, "Support vector method for novelty detection," *Adv. Neural Inf. Process. Syst.*, no. January, pp. 582–588, 2000.

[13] F. Tony Liu, K. Ming Ting, and Z.-H. Zhou, "Isolation Forest ICDM08," *Icdm*, 2008, [Online]. Available: https://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf%0Ahttps://cs.nju.edu.cn/zhouzh/zhouzh.files/publication/icdm08b.pdf?q=isolation-forest.

[14] L. Ruff *et al.*, "Deep one-class classification," *35th Int. Conf. Mach. Learn. ICML 2018*, vol. 10, pp. 6981–6996, 2018.

[15] S. Liang, Y. Li, and R. Srikant, "Enhancing the reliability of out-of-distribution image detection in neural networks," *6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc.*, pp. 1–15, 2018.

[16] I. Golan and R. El-Yaniv, "Deep anomaly detection using geometric transformations," *Adv. Neural Inf. Process. Syst.*, vol. 2018-Decem, no. NeurIPS, pp. 9758–9769, 2018.

[17] Z. Yang, I. S. Bozchalooi, and E. Darve, "Regularized Cycle Consistent Generative Adversarial Network for Anomaly Detection."

[18] D. T. Nguyen, Z. Lou, M. Klar, and T. Brox, "Anomaly detection with multiple-hypotheses predictions," *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 8418–8432, 2019.

[19] S. Pidhorskyi, R. Almohsen, D. A. Adjeroh, and G. Doretto, "Generative probabilistic novelty detection with adversarial autoencoders," *Adv. Neural Inf. Process. Syst.*, vol. 2018-Decem, no. Nips, pp. 6822–6833, 2018.

[20] S. Zhai, Y. Cheng, W. Lu, and Z. Zhang, "Deep structured energy based models for anomaly detection," *33rd Int. Conf. Mach. Learn. ICML 2016*, vol. 3, pp. 1742–1751, 2016.

[21] T. Schlegl, P. Seeböck, S. M. Waldstein, G. Langs, and U. Schmidt-Erfurth, "f-AnoGAN: Fast unsupervised anomaly detection with generative adversarial networks," *Med. Image Anal.*, vol. 54, pp. 30–44, 2019, doi: 10.1016/j.media.2019.01.010.

[22] H. Zenati, C. S. Foo, B. Lecouat, G. Manek, and V. R. Chandrasekhar, "Efficient GAN-Based Anomaly Detection," 2018, [Online]. Available: http://arxiv.org/abs/1802.06222.

[23] C. Li *et al.*, "ALICE : Towards Understanding Adversarial Learning for Joint Distribution Matching arXiv : 1709 . 01215v2 [stat . ML] 5 Nov 2017," no. Nips, pp. 1–22, 2017.

[24] A. Makhzani and B. Frey, "Winner-take-all autoencoders," *Adv. Neural Inf. Process. Syst.*, vol. 2015-Janua, pp. 2791–2799, 2015.

[25] V. Dumoulin *et al.*, "Adversarially learned inference," *5th Int. Conf. Learn. Represent. ICLR 2017 - Conf. Track Proc.*, pp. 1–18, 2017.

[26] B. Zong *et al.*, "Deep autoencoding Gaussian mixture model for unsupervised anomaly detection," *6th Int. Conf. Learn. Represent. ICLR 2018 - Conf. Track Proc.*, pp. 1–19, 2018.

[27] T. Salimans, I. Goodfellow, V. Cheung, A. Radford, and X. Chen, "Improved Techniques for Training GANs," pp. 1–10.