«نشریه علمی (فصلنامه) «انرژی ایران
دوره ۲۸، شماره۱، بهار ۱۴۰۴، صفحه ۶۵-۸۱

رویکرد مبتنی بر یادگیری عمیق برای شناسایی حمله تزریق داده جعلی در شبکه‌های هوشمند

علیرضا سلطانی
کارشناسی ارشد برق، دانشکده مهندسی برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، ایران
alireza_soltani@modares.ac.ir
دکتر حمیدرضا بقایی‌کاشی
استادیار گروه برق، دانشکده مهندسی برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، ایران (نویسنده مسئول)
hrbaghaee@modares.ac.ir
دکتر محمودرضا حقی فام
استاد گروه برق، دانشکده مهندسی برق و کامپیوتر، دانشگاه تربیت مدرس، تهران، ایران
haghifam@modares.ac.ir

تاریخ دریافت:
۱۴۰۴/۰۹/۱۲
تاریخ پذیرش:
۱۴۰۴/۱۱/۰۶

کلمات کلیدی:
شبکه هوشمند
تشخیص ناهنجاری
حمله سایبری
یادگیری عمیق

چکیده

با گسترش شبکه‌های توزیع هوشمند و افزایش تهدیدات سایبری، ایمن‌سازی سیستم‌های قدرت به ضرورتی حیاتی تبدیل شده است. حملات تزریق داده جعلی از جمله مخرب‌ترین تهدیدات هستند که با دور زدن مکانیزم‌های تشخیص سنتی، ثبات سیستم را به خطر می‌اندازند. در این پژوهش، یک مدل تشخیصی مبتنی بر **Graph Autoencoder** ارائه شده است که با رویکردی پیشگیرانه به شناسایی این حملات در شبکه توزیع ۳۳ IEEE باسه می‌پردازد. مدل پیشنهادی در چارچوب یک معماری یادگیری عمیق طراحی گردیده و با بهره‌گیری از قابلیت‌های شبکه عصبی گرافی، الگوهای عادی عملکرد شبکه را آموخته و از طریق محاسبه خطای بازسازی، حملات را شناسایی می‌نماید. نتایج حاکی از آن است که مدل ارائه‌شده بالای ۹۸ درصد قادر به تشخیص مؤثر ناهنجاری‌ها بوده و منجر به بهبود عملکرد سیستم حفاظتی می‌گردد.

۱.مقدمه

شبکه قدرت یکی از زیرساخت های بنیادین اقتصادی – اجتماعی است که سلامت، پایداری و دسترسی پیوسته آن برای عملکرد جوامع مدرن حیاتی است. در سال‌های اخیر، تحول فناوری اینترنت صنعتی اشیاء (IIoT[۱]) و ظهور شبکه‌های هوشمند[۲] موجب ادغام گسترده فناوری اطلاعات و ارتباطات با شبکه قدرت شده است؛ این ادغام به‌طور چشمگیری کارآمدی و قابلیت اطمینان شبکه را افزایش داده است ولی هم‌زمان چالش‌های امنیتی و سایبری نوینی نیز ایجاد کرده است[۱].

یکی از ستون‌های عملیاتی مرکز کنترل سیستم قدرت، فرآیند تخمین حالت[۳] است که از داده‌های خام گردآوری‌شده توسط سیستم‌های SCADA[۴] وضعیت سیستم را محاسبه می‌کند و نتایج آن ورودی بسیاری از تابع‌های اجرایی مانند پخش بار اقتصادی و تحلیل پایداری است. از این رو، هرگونه دست‌کاری یا اخلال در نتایج برآورد حالت می‌تواند به خطاهای تصمیم‌گیریِ تزریق بار نامناسب، یا در نهایت وقوع رویدادهای بحرانی (مانند تلفات و قطعی) منجر شود[۲].

حملات FDI[۵]

حملات تزریق داده جعلی (FDI) نوعی حمله‌ای است که به‌طور خاص بر سامانه برآورد حالت هدف‌گیری می‌کند: مهاجم با تغییر مقادیر اندازه‌گیریِ ارسالی از سوی شمارنده‌ها یا PMU[۶]ها می‌تواند مقدارهای ورودی تخمین‌گر حالت را به‌گونه‌ای دست‌کاری کند که خطای ساختاری در برآورد حالت ایجاد شود، در حالی که مکانیزم‌های متداول تشخیص داده‌های بد (BDD[۷]) را دور می‌زند. کارهای ابتدایی نشان دادند که با داشتن دانشی کافی از مدل شبکه، مهاجم می‌تواند حمله‌ای کاملاً پنهانی[۸] بسازد که با ساختار ریاضی برآورد حالت سازگار باشد و به همین دلیل به‌سختی قابل‌کشف است. بعداً نشان داده شد حتی با دسترسی ناقص یا دانش جزئی درباره پیکربندی شبکه، حملات قابل‌تأثیر و تا حدی غیرقابل‌کشف قابل طراحی هستند و بنابراین دفاع صرفاً مبتنی بر BDD ناکافی است [۳].

در این مقاله با استفاده روش های مبتی بر هوش مصنوعی برای پیشگیری و تشخیص حملات تزریق داده جعلی در شبکه نمونه (IEEE ۳۳ bus) می‌پردازیم.

مقایسه با روش های ارائه شده در مراجع:

در [٤] یک حمله‌ی FDI جعبه‌سیاه را ارائه می‌دهد که در آن بردار حمله با استفاده از یک شبکه‌ی مولد تخاصمی (GAN) تولید می‌شود تا اختلال‌هایی مشابه داده‌های عادی سیستم ایجاد کند. سپس، این بردار حمله از طریق دستکاری ماژول های اندازه‌گیری شبکه به سیستم تزریق می‌شود. این حمله بلافاصله پس از وقوع یک خطا در سیستم اجرا می‌شود تا هم پنهان‌ماندن از مکانیزم‌های تشخیص داده بد افزایش یابد و هم

[۱] Industrial Internet of Things
[۲] Smart grids
[۳] State estimation
[٤] Supervisory Control And Data Acquisition
[٥] False Data Injection Attack
[٦] Phasor Measurement Unit
[۷] Bad Data Detection
[۸] Stealthy

اثر تخریبی آن بیشینه شود، در [۵]، [۶]، [۷] و [۸] با استفاده از دانش ساختار شبکه(ماتریس ژاکوبین) حمله پس از دور زدن BDD ساخته شده است.(تخمین‌گر حالت حمله را تشخیص نمی‌دهد.) در جدول۱ این منابع از نظر دقت تشخیص حملات از داده های سالم بررسی می‌شوند.

مدل	شبکه	دقت	فراخوانی	امتیاز F۱
AlexNet[٤]	IEEE ٥٧-bus	٩٨٫٤٠%	-	-
AlexNet[٤]	IEEE ١٤٥-bus	٩٧,١٠٪	-	-
SFTCN[۹][٥]	IEEE ١٤-bus	٩٩,٦٨٪	٩٨,٨٣٪	٩٩,٣٦٪
SFTCN[٥]	IEEE ١١٨-bus	٩٧,٢٠٪	٩١,٨٥٪	٩٤,٢٦٪
A-BiTG[۱۰][٦]	IEEE ١٤-bus	٩٥,٤٧٪	٩٧,٢٧٪	٩٦,٣٧٪
GMM[۱۱][٧]	IEEE ١٤-bus	٩٥٪<	-	-
LSTM[۱۲][٨]	IEEE ٣٠-bus	٩٤,٢٤٪	-	-
LSTM[٨]	IEEE ١٤-bus	٩٤,٥٢٪	-	-

جدول(۱).مقایسه دقت تشخیص روش های مختلف

در روش پیشنهادی در این پروژه بدون دانستن مشخصات شبکه(ماتریس ژاکوبین) با استفاده از روش زیرفضا حمله طراحی شده است و حمله قادر به عبور از BDD مرسوم شبکه است، از مدل تست های متعدد گرفته شده است و دقت تشخیص حمله تقریبا بالاتر از ۹۹ درصد است و حتی اگر یکی از اندازه گیری ها مورد حمله باشد، مدل قادر به تشخیص آن است.

۲.مبانی نظری

در اغلب سیستم های قدرت، وضعیت سیستم با یک بردار حالت x (مثلاً بردار ولتاژها و زاویه‌ها) مدل می‌شود و سنجش‌ها (اندازه گیری ها) z توسط رابطه ۱ نمایش داده می‌شود.

رابطه ۱: $$z = H(x) + e$$

که H و e به ترتیب نشان دهنده ماتریس ژاکوبین شبکه و بردار خطای اندازه‌گیری تصادفی است که معمولاً به عنوان توزیع گاوسی مستقل با میانگین صفر(طبق رابطه ۲) فرض می‌شود[۱].

رابطه ۲: $$\mathcal{N}(\cdot, \sigma^2 I) \sim e$$

حمله تزریق داده جعلی را می‌توان به‌صورت افزودن یک بردار حمله a به اندازه‌گیری‌ها مدل کرد:

رابطه ۳: $$H(x) + e + a = a + z = \acute{z}$$

که $\acute{z}$ همان بردار حاصل از حمله است.

[۹] Spatial Feature-based Temporal Convolutional Network
[۱۰] Attention-based Bidirectional Temporal Gating
[۱۱] Gaussian Mixture Model
[۱۲] Long-Short Term Memory

هدف حمله‌ی تزریق داده‌ی جعلی گمراه کردن اپراتور سیستم است تا وضعیت تخمینی و مخدوش‌شده $x = x + c$ را به‌عنوان یک تخمین معتبر در نظر بگیرد، که در آن $c \neq 0$ انحراف حالت سیستم قدرت است. برای دستیابی به این هدف، مهاجم مقادیر اندازه‌گیری دریافتی در مرکز کنترل را به $\acute{z} = z + a$ تغییر می‌دهد[۱].

روش‌های متعارف تشخیص ناهنجاری مبتنی بر باقیمانده، باقیمانده اندازه‌گیری(R) را با یک آستانه از پیش تعیین شده (τ) مقایسه می‌کنند تا بررسی کنند که آیا اندازه‌گیری‌های بی‌کیفیت (یعنی خراب یا مخدوش شده) در سیستم وجود دارند یا خیر[۹].در این حالت، آشکارساز[۱۳] به محض اینکه شرط زیر(رابطه ۴) برقرار شود، وجود یک حمله را اعلام می‌کند:

رابطه ۴: $R = z - Hx > \tau$

برای دور زدن مکانیسم BDD، بردار حمله باید ساختاریافته باشد، به‌گونه‌ای که $a = Hc$. در چنین مواردی R (باقیمانده[۱۴]) در رابطه ۴ بدون تغییر می‌ماند:

رابطه ۵: $\|z + a - H(x + c)\| = \|z - Hx\|$

و در نتیجه(طبق رابطه ۵) حمله می‌تواند از BDD عبور کند. بر این اساس، اپراتور سیستم قدرت به اشتباه $x + c$ را یک تخمین معتبر در نظر می‌گیرد و بنابراین یک بردار خطای c معرفی می‌شود.

حمله زیر فضا[۱۵]:

حملات زیرفضا مبتنی بر این ایده هستند که مهاجم می‌تواند بدون دانش کامل از پارامترهای سیستم، تنها با تحلیل داده‌های تاریخی، ساختار شبکه را یاد گرفته و حملات غیرقابل تشخیص طراحی کند[۱۰].

فرض شود که U یک ماتریس پایه برای فضای ستونی $R(H)$ باشد و U_1 زیر ماتریس از U است که با حذف سطر های مربوط به سنسور های مورد تهاجم (S_A) به‌دست می‌آید.

شرط امکان‌پذیری حمله غیرقابل مشاهده:

یک حمله غیرقابل مشاهده امکان‌پذیر است اگر و تنها اگر U_1 رتبه ستونی کامل نداشته باشد.

اگر U_1 رتبه کامل داشته باشد، آنگاه تنها بردار v که در $U_1 v = 0$ صدق کند، بردار صفر است.

اگر U_1 رتبه کامل نداشته باشد، آنگاه فضای پوچ $N(U_1)$ دارای بعد مثبت است[۱۰].

$N(U_1)$ همان فضای پوچ ماتریس U_1 است؛ یعنی مجموعه‌ای از تمام بردارهایی که اگر در U_1 ضرب شوند، حاصل بردار صفر می‌شود، اگر این فضا بعد مثبت داشته باشد، امکان ساخت حمله غیرقابل تشخیص وجود دارد.

[۱۳] Detector
[۱۴] Residual
[۱۵] Subspace Attack

الگوریتم طراحی حمله:

$\{z_1, z_2, \dots z_k\}$ مقادیر بردار اندازه‌گیری های گذشته هستند که مهاجم از آن ها استفاده می‌کند و مجموعه سنسور هایی که مهاجم برای حمله انتخاب می‌کند، S_A هستند.

برای مرکز سازی داده ها:

رابطه ۶: $\bar{z} = \frac{1}{k}\sum_{i=1}^{k} z_i$

رابطه ۷: $z_c^{(i)} = z_i - \bar{z}$

که $z_c^{(i)}$ بردار مرکزسازی شده اندازه‌گیری ها و $\bar{z}$ بردار میانگین داد ها هستند، برای ساده سازی فرض می‌کنیم بردار های نویز $e_1, \dots, e_k$ مستقل و دارای توزیع مشابه[16]هستند، بردار های حالت $x_1, \dots, x_k$ نیز وضعیت مشابهی مانند بردار های نویز دارند و دارای ماتریس کوواریانس معین مثبت Σ_x هستند، بردارهای نویز و بردارهای حالت نیز ناهمبسته هستند. در این صورت، ماتریس کوواریانس z به‌صورت زیر محاسبه می‌شود:

رابطه ۸: $\Sigma_z \triangleq E[(z_i - E[z_i])(z_i - E[z_i])^T] = H\Sigma_x H^T + \sigma^2 I$

با توجه به اینکه $\widehat{\Sigma_z}$ ماتریس مثبت معین است در نتیجه معکوس پذیر است و چون H^T نیز رتبه کامل سطری دارد، ضرب آن در سمت راست رتبه ستونی را تغییر نمی‌دهد پس:

رابطه ۹: $\mathcal{R}\left(H\hat{\Sigma}_x H^T\right) = \mathcal{R}(H)$

طبق رابطه ۹ زیر فضای پوششی$(\mathcal{R})$ در H معادل زیر فضای پوششی $H\hat{\Sigma}_x H^T$ است.

بنابراین، در عمل می‌توانیم یک ماتریس پایه برای $\mathcal{R}(H)$ را با اعمال تجزیه مقدار منفرد روی ماتریس کوواریانس نمونه ای $\hat{\Sigma}_z$ تخمین بزنیم:

رابطه ۱۰: $\hat{\Sigma}_z = \frac{1}{k-1}\sum_{i=1}^{k} z_c^{(i)} \left(z_c^{(i)}\right)^T$

که k تعداد نمونه ها و T نشان‌دهنده ترانهاده ماتریس هستند.

محاسبه مقادیر منفرد(SVD)[17]:

رابطه ۱۱: $\hat{\Sigma}_z = U\boldsymbol{\Lambda}V^T$

که U ماتریس بردار های ویژه چپ، $\boldsymbol{\Lambda}$ ماتریس مقادیر ویژه و V^T ترانهاده ماتریس بردارهای ویژه راست هستند.

[16] i.i.d: Independent And Identically Distributed
[17] Singular Value Decomposition

ماتریس $\widehat{U}_1$ حاصل از حذف سطر های مربوط به S_A در U است، مقادیر منفرد این ماتریس به صورت زیر است:

رابطه ۱۲:
$$\widehat{U}_1 = \widetilde{U}\widetilde{\Lambda}\widetilde{V}^T$$

اگر حمله امکان‌پذیر باشد، برای هر بردار غیرصفر $v \in N(\widehat{U}_1)$ بردار $a = Uv$ یک بردار حمله غیر قابل مشاهده است[۱۰].

این بردار حمله در هر ردیف از $a \neq 0$ به داده های واقعی شبکه(در اینجا داده‌های ولتاژ) اضافه می‌شود و اندازه‌گیری آن سنسور آلوده(S_A) می‌شود.

این حمله به‌گونه‌ای طراحی شده که فقط داده سنسورهای حمله‌شده را تغییر دهد ولی در محاسبات باقیمانده سیستم تشخیص سنتی غیرقابل مشاهده باقی بماند، مدل شبکه عصبی ارائه شده با یادگیری وابستگی‌های پیچیده‌تر شبکه، می‌تواند این حمله هوشمند را از طریق خطای بازسازی بالا شناسایی کند.

۳.روش تحقیق

هدف روش، طراحی و ارزیابی یک چارچوب داده‌محور برای شناسایی و مکان‌یابی حمله سایبری در سامانه اندازه‌گیری شبکه توزیع، حملات تزریق داده جعلی است. رویکرد مقاله مبتنی بر یک مدل یادگیرنده گرافی است که ورودی‌های گره‌ای را بازسازی می‌کند و از خطاهای بازسازی برای برچسب‌گذاری نمونه‌ها و شناخت ناهنجاری‌ها استفاده می‌نماید. این بخش روند کار آزمایشی و روش‌شناختی را تشریح می‌کند.[۱۱]

برای مدل‌سازی و شناسایی ناهنجاری‌های ناشی از حملات تزریق داده جعلی از چارچوب GAE[۱۸] مبتنی بر GCN [۱۹]استفاده شد. در این طراحی، یک رمزگذار جهت یادگیری نمایش نهفته[۲۰] هر گره و یک رمزگشا[۲۱] مبتنی بر GCN جهت بازسازی ویژگی‌های گره‌ها به‌کار رفته است. ایده پایه این است که شبکه پس از آموزش روی داده‌های سالم، الگوهای همبستگی فضایی بین باس‌ها (گره‌ها) و ویژگی‌های آن‌ها را یاد می‌گیرد؛ در صورت وقوع حمله، خطای بازسازی گره(ها) افزایش می‌یابد که می‌تواند به عنوان شاخص تشخیص و مکان‌یابی مورد استفاده قرار گیرد[۱۲]، [۱۳].

ماتریس مجاورت A: در شبکه اتصال بین باس ها اگر بین گره(باس) iوj اتصال فیزیکی(خط) وجود داشته باشد; $A_{i,j} = 1$ و در غیر این‌صورت $A_{i,j} = 0$ خواهد شد.

ماتریس خود همراه $\hat{A}$: طبق رابطه ۱۳ افزودن ماتریس همانی(I) یعنی هر گره خودش نیز در میانگین گیری همسایه ها لحاظ می‌شود، این عمل اجازه می‌دهد اطلاعات خود گره به محاسبه ویژگی جدید اضافه شود.

رابطه ۱۳:
$$\hat{A} = A + I$$

[۱۸] Graph Autoencoder
[۱۹] Graph Convolutional Network
[۲۰] embedding
[۲۱] decoder

رابطه ۱۴: $\widehat{D}_{ii} = \sum_j \hat{A}_{ij}$

که $\widehat{D}$ جمع سطرهای $\hat{A}$ ؛ نشان‌دهنده تعداد همسایگان هر گره است.

ماتریس نرمال شده $\hat{A}$: طبق رابطه ۱۵ نرمال سازی $\hat{A}$ باعث می‌شود که از ناپایداری مقیاس جلوگیری شود(از تسلط گره های با درجه بالا در انتشار پیام جلوگیری می‌کند.) و در بسیاری از پیاده سازی های GCN این فرم را مبنای لایه کانولوشن قرار می‌دهند.

رابطه ۱۵: $\hat{A} = \widehat{D}^{-۱/۲}\hat{A}\widehat{D}^{-۱/۲}$

لایه پایه GCN: هر لایه GCN ویژگی گره هارا با میانگین‌گیری وزنی از همسایگان و تبدیل خطی $W^{(l)}$ ترکیب می‌کند، سپس تابع غیرخطی σ اعمال می‌شود.

رابطه ۱۶: $H^{(l+۱)} = \sigma\big(\hat{A}\,H^{(l)}W^{(l)} + b^{(l)}\big), \qquad H^{(۰)} = X$

که در رابطه ۱۶، $H^{(l)}$; نشان دهنده خروجی لایه l، $W^{(l)}$; ماتریس وزن لایه l، $b^{(l)}$; ماتریس بایاس[۲۲]، σ; تابع فعال سازی[۲۳]، X; ماتریس ویژگی ورودی، هستند[۱۴].

در این پژوهش از تابع فعال ساز ReLU برای لایه ها استفاده شده است.

پیاده‌سازی ارائه‌شده شامل سه کلاس کلیدی است:

GCNEncoder

این مؤلفه مسئول یادگیری نمایش فشرده‌شده گره‌ها در فضای نهفته است. معماری آن شامل دو لایه گراف کانولوشنال (GCN) متوالی است که ورودی اولیه ماتریس ویژگی‌های گره‌ای X و ماتریس مجاورت A را دریافت کرده و بردارهای نهفته Z را برای هر گره تولید می‌کند.

GCNDecoder

این بخش مسئول بازسازی ویژگی‌های اولیه از بردارهای نهفته است. در این پیاده‌سازی، رمزگشا از دو لایه گراف کانولوشنال معکوس[۲۴]تشکیل شده که بردارهای نهفته Z را به مقادیر بازسازی‌شده ولتاژ $\hat{X}$ نگاشت می‌کند.

GraphAutoEncoder

این کلاس با ترکیب رمزگذار و رمزگشا، مدل کامل را تشکیل می‌دهد، خروجی‌های این مدل شامل بردارهای نهفته Z و مقادیر بازسازی‌شده ولتاژ $\hat{X}$ است[۱۳], [۱۵].

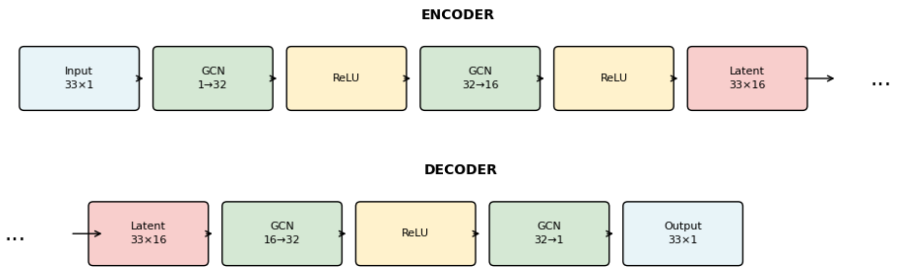
[۲۲] Bias
[۲۳] Activation Function
[۲۴] Deconvolution

تابع هزینه[۲۵]:

با فرض ثابت بودن توپولوژی شبکه در شرایط عادی، در پیاده‌سازی فعلی تنها بازسازی ویژگی‌ها (ولتاژها) انجام می‌شود؛ بنابراین تابع هزینه اصلی میانگین مربعات خطا [۲۶]بین X و $\hat{X}$ است:

رابطه ۱۷: $$\mathcal{L}_X = |X - \hat{X}|_F^2 = \sum_{i=1}^{N} |x_i - \hat{x}_i|_2^2$$

که در رابطه۱۷ x_i نشان‌دهنده مقادیر واقعی ولتاژ ها، $\hat{x}_i$ مقادیر ولتاژ های ساخته شده توسط مدل و $\mathcal{L}_X$ تابع هزینه هستند.



شکل(۱)، نمودار شبکه عصبی GAE پروژه

شکل ۱، نمودار شبکه عصبی به کار رفته در این پروژه را نشان می‌دهد که، Encoder از دو لایه GCN با توابع فعال‌ساز ReLU تشکیل شده که ابعاد ویژگی‌ها را از ۱ به ۳۲ و سپس به ۱۶ کاهش می‌دهد ،Decoder نیز شامل دو لایه GCN است که فرآیند معکوس را برای بازسازی ویژگی‌های اولیه انجام می‌دهد.

(نمایش) نهفته[۲۷](z):

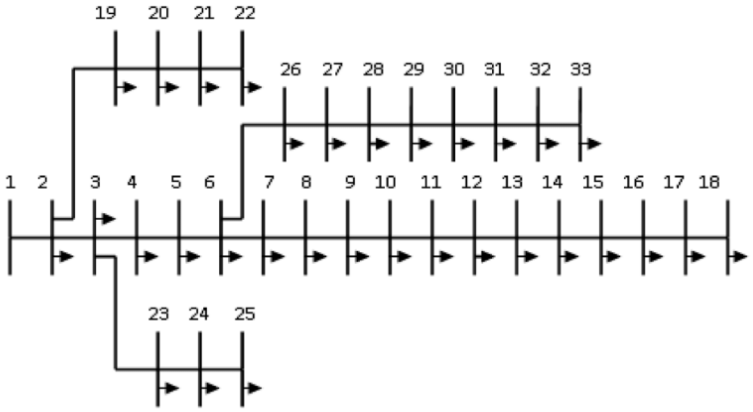
بردارِ نهفته‌ی هر گره است که توسط Encoder تولید می‌شود خروجی z با ابعاد، ابعاد بردار نهفته در تعداد نود ها(در پیاده‌سازی ارائه شده ۳۳×۱۶)این بردار بیانگر فشرده‌سازی اطلاعات محلی هر باس بوده و از طریق تجمیع ویژگی‌های همسایگان در لایه‌های GCN، مؤلفه‌های ساختاری و ویژگی‌محور شبکه را کدگذاری می‌کند، Decoder همین نمایش نهفته را برای بازسازی ولتاژ گره‌ها به‌کار می‌گیرد، فرآیند آموزش با کمینه‌سازی خطای بازسازی، به مدل امکان می‌دهد تا یک نمایش فشرده و معنادار از وضعیت عادی شبکه را فرا گیرد.

[۲۵] Loss Function
[۲۶] MSE: Mean Square Error
[۲۷] Latent

در این پروژه برای تشخیص حملات، این شبکه ابتدا الگوهای عادی عملکرد شبکه را از داده‌های آموزشی یاد می‌گیرد. هنگامی که داده‌های حمله به مدل ارائه می‌شوند، به دلیل انحراف از الگوهای عادی آموخته شده، خطای بازسازی مدل به طور قابل توجهی افزایش می‌یابد. این خطای بالا به عنوان شاخصی برای تشخیص حملات مورد استفاده قرار می‌گیرد[۱۶].

۴.توصیف داده‌ها

مجموعه داده مورد استفاده در این پژوهش شامل مجموعه‌ای از سناریوهای شبیه‌سازی‌شده عملیاتی از یک شبکه توزیع نمونه (شبکه IEEE ۳۳ باسه) است.

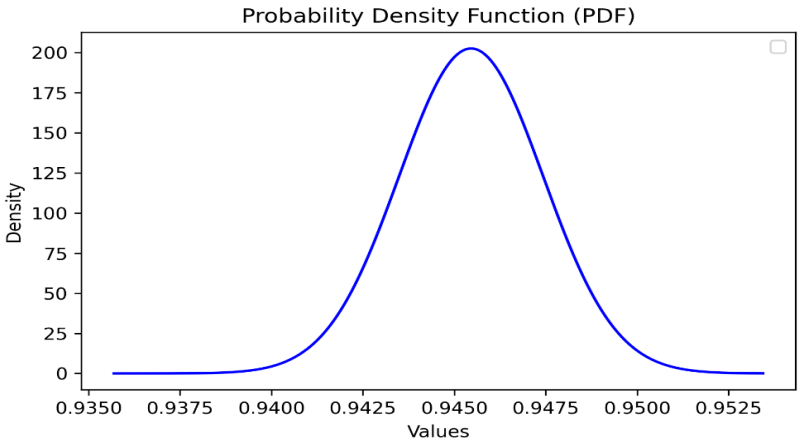


شکل(۲). شبکه ۳۳ باسه IEEE

شکل۲ شبکه ۳۳ باسه IEEE و نحوه اتصال باس ها به‌یک‌دیگر و بار های شبکه را نشان می‌دهد. تمامی پارامتر های این شبکه (بار، امپدانس خطوط، طول خطوط و...) طبق استاندارد IEEE در نرم افزار دیگ سایلنت پیاده‌سازی شده‌اند.

هر سناریو نمایانگر وضعیت شبکه در یک زمان ثابت است و شامل مشاهدات گره‌ای و یالی استاندارد می‌باشد. نحوه بدست آمدن سناریو ها از تغییر بار های شبکه به صورت توزیع نرمال در بار مرجع شبکه IEEE است(بار ها با ضریب ۰,۶ تا ۱,۱ برابر بار اصلی شبکه هستند و توزیع نرمال با میانگین ۰,۸۵ و انحراف از معیار ۰,۱۲۵ دارند.) داده‌ها به‌صورت مجموعه‌ای از فایل‌های جداسازی‌شده نگهداری می‌شوند و برای ارزیابی مدل به مجموعه‌های آموزش، اعتبارسنجی و تست تقسیم شده‌اند. در این پژوهش از ولتاژ های شبکه و در ۶۰۰۰۰ سناریو نمونه برداری شده است، چون شبکه ۳۳ باسه است، بردار ورودی شبکه به صورت ۶۰۰۰۰×۳۳ است، و داده ها بدون تغییر از پخش بار بدست می‌آیند.(به طوری که هر سطر نمایانگر یک سناریو و هر ستون مربوط به ولتاژ یک باس خاص می‌باشد.) داده ها در نرم افزار اکسل و به صورت per-unit شده ذخیره شده‌اند.

در شکل۳ تابع چگالی احتمال ولتاژ های بدست آمده از سناریو ها مشاهده می‌شود، این تابع چگالی احتمال مربوط به باس ۹ شبکه تست می‌باشد.



شکل(۳). تابع چگالی احتمال ولتاژ باس۹

نمونه‌های آلوده به‌صورت کنترل‌شده تولید شدند: برای هر نمونه چند سنسور (گره) به‌طور تصادفی انتخاب و اختلال جمع‌شونده‌ای(طبق روش زیر فضا) با دامنه‌ای منطقی به اندازه‌گیری‌ها اضافه شد تا مقادیر در چارچوب فیزیکی باقی بمانند. (توسط سیستم های سنتی تشخیص خطا یا داده آلوده تشخیص داده نشوند.) الگوهای متنوعی و با شدت‌های مختلف ساخته شد و داده های تولید شده به عنوان داده های آلوده به حمله تزریق داده جعلی در تست مدل استفاده شدند، حداکثر دامنه حمله اعمال شده ۰,۱ پریونیت به داده های ولتاژ است، زمان شروع حمله به شرایط خاصی بستگی ندارد(مثلا در حین خطا، قبل و بعد از خطا و...) و در هر زمانی قابل اعمال است.

۵.نتایج اجرای مدل

برای تشخیص ناهنجاری های ناشی از حملات تزریق داده جعلی، مدل اصلی مورد استفاده یک Graph Autoencoder است که ورودی‌های گره‌ای و یالی را به فضای پنهان نگاشت و سپس با یک decoder بازسازی مقادیر اصلی را انجام می‌دهد. خطای بازسازی در سطح باس ها به‌عنوان معیار تشخیص حمله استفاده شده است. در این پژوهش ساختار گراف و اتصالات شبکه تست برای شبکه عصبی تعریف و معین شده است. شبکه عصبی با شناخت ساختار و رابطه داده های سناریو های آموزش، آموزش می‌بیند که در مواجهه با سناریو های جدید کدام باس ها سالم و کدام باس ها مورد حمله تزریق داده جعلی قرار گرفته‌اند.Graph Autoencoder به‌دلیل توانایی در یادگیری بازنمایی‌های موثر از ساختار و داده‌های شبکه قدرت، و مکانیسم تشخیص ناهنجاری بر اساس خطای بازسازی، گزینه مناسبی برای تشخیص حملات تزریق داده جعلی است.) در این شبکه از مدل بدون نظارت [۲۸]شبکه عصبی استفاده شده است که برای آموزش این شبکه فقط از داده های سالم و نرمال شبکه که با تغییر بار های شبکه بدست آمده‌اند استفاده شده است، برای آموزش مدل در حدود ۸۵ درصد داده ها برای آموزش و ۱۵ درصد آن ها به اضافه قسمت حمله برای تست مدل استفاده شدند، البته برای تست بیشتر داده های بیشتری علاوه بر داده های اصلی درنظر گرفته‌شده‌اند که در ادامه به آن می‌پردازیم.

[۲۸] Unsupervise

داده‌های سالم در یک زیرمجموعه محدود از فضای ورودی یا زیرفضا قرار می‌گیرند و GAE پس از آموزش، توانایی بازسازی دقیق نمونه‌هایی را دارد که در نزدیکی این زیرفضا قرار دارند. حملات تزریق داده جعلی، به‌طور کلی، مقادیر ورودی را به بیرون از آن زیرفضا می‌برند (یا آن‌گونه تغییر می‌دهند که نمونه دیگر در توزیع واقعی داده‌های سالم نیست). بنابراین زمانی که یک نمونه آلوده به مدل اعمال می‌شود، decoder قادر به بازسازی دقیق آن نخواهد بود و خطای بازسازی (به ویژه در گره‌های تحت تاثیر حمله) به‌صورت معناداری بزرگ می‌شود(از آستانه بیشتر می‌شود). از این اختلاف[۲۹] می‌توان برای تشخیص نمونه آلوده استفاده نمود.

تعیین آستانه:

برای تشخیص اینکه یک نمونه (سناریو) ناهنجار است یا خیر، از خطای بازسازی نمونه‌ای (طبق روابط ۱۸ و ۱۹) استفاده می‌کنیم:

رابطه ۱۸:
$$\mathrm{MSE}^{(s)} = \frac{1}{N}\sum_{i=1}^{N} |x_i^{(s)} - \widehat{x_i^{(s)}}|_2^2$$

رابطه ۱۹:
$$\tau = Q_p\left(\{\mathrm{MSE}^{(v)}\}_{v\in\mathcal{V}}\right)$$

آستانه تصمیم τ بر مبنای صدک امپریک [۳۰] p از توزیع $\mathrm{MSE}^{(s)}$ روی مجموعه ارزیابی بدون حمله(سالم) $\mathcal{V}$ تعیین می‌شود، قاعده تشخیص به این صورت است که، نمونه s را ناهنجار فرض می‌کنیم اگر رابطه ۲۰ برقرار باشد[۱۷].

رابطه ۲۰:
$$\mathrm{MSE}^{(s)} > \tau$$

در این پیاده سازی مقدار $p = ۰.۹۵$ در نظر گرفته شد.

شبکه عصبی با batch_size برابر ۱۲۸ و تعداد epochs برابر ۱۰۰ اجرا شد و بهترین نتیجه مدل با train_loss برابر ۰,۰۰۰۰۰۶۷۷۴۸ ذخیره شد.

برای اینکه داده های بدست آمده قابل اعتماد باشند، مدل در سه seed[۰،۱،٤۲] امتحان شده است.

این شبکه عصبی بارها برای انواع سناریو های حملات تزریق داده جعلی تست شد، روابط و نتایج پارامتر های سنجش عملکرد شبکه به صورت زیر ارائه می‌گردد:

رابطه ۲۱:
$$Precision^{[31]} = TP^{[32]} / (TP + FP^{[33]})$$

[۲۹] Reconstruction Residual
[۳۰] Empirical Percentile
[۳۱] دقت
[۳۲] True Positive
[۳۳] False Positive

رابطه ۲۲: $Recall^{۳۴} = TP / (TP + FN^{۳۵})$

رابطه ۲۳: $F1 - Score = ۲ \times (Precision \times Recall) / (Precision + Recall)$

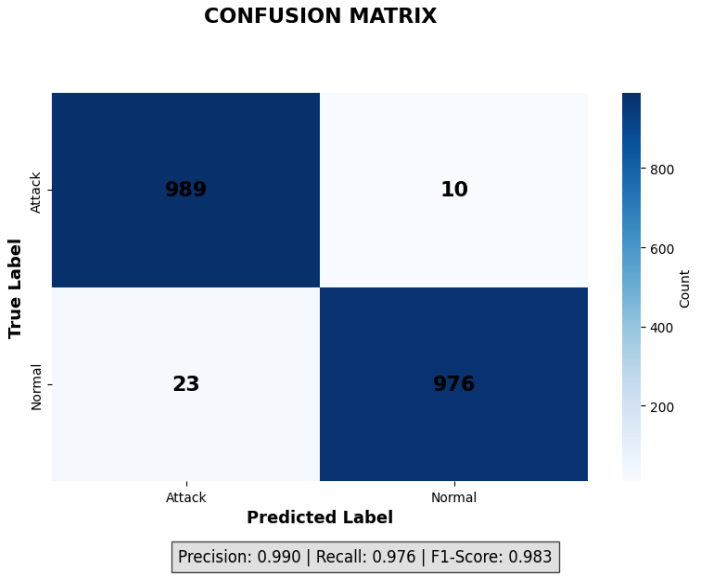
رابطه۲۱، دقت: نسبت نمونه‌های مثبت صحیح به تمام نمونه‌های پیش‌بینی‌شده به عنوان مثبت.

رابطه۲۲، حساسیت(فراخوانی): نسبت نمونه‌های مثبت صحیح به تمام نمونه‌های مثبت واقعی.

جدول(۲). معیار های بررسی مدل

seed	دقت	فراخوانی	امتیاز F۱
۰	۹۹,۳٪	۹۷,۵٪	۹۸,٤٪
۱	۹۸,٦٪	۹۷,۳٪	۹۷,۹٪
٤۲	۹۹,۰٪	۹۸,۱٪	۹۸,۵٪
میانگین	۹۸,۹٪	۹۷,٦٪	۹۸,۳٪

در شکل۴ ماتریس درهم‌ریختگی[۳۶](میانگین) نمایش داده شده است، که نمایان گر تعداد حملات و توانایی مدل در تشخیص حملات و داده های نرمال از یکدیگر است.



شکل(۴). ماتریس درهم‌ریختگی مدل

[۳۴] فراخوانی
[۳۵] False Negative
[۳۶] Confusion matrix

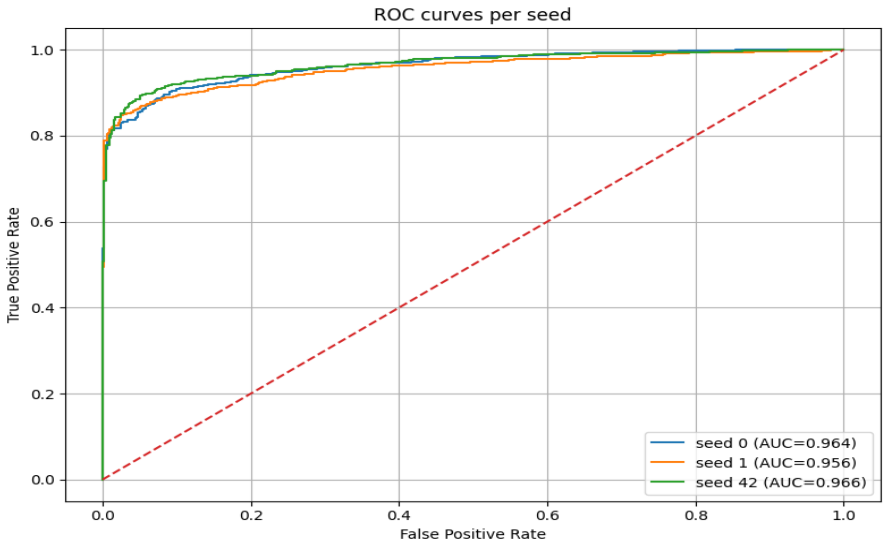
معیار دیگر برای سنجش توانایی مدل در تشخیص حملات، منحنی عملکرد سیستم [۳۷]است.

رابطه ۲۴: $$FPR = FP/(FP + TN)$$

رابطه ۲۵: $$\text{AUC-ROC}^{[38]} = \int_{0}^{1} TPR(f)\ d\,(FPR(f))$$

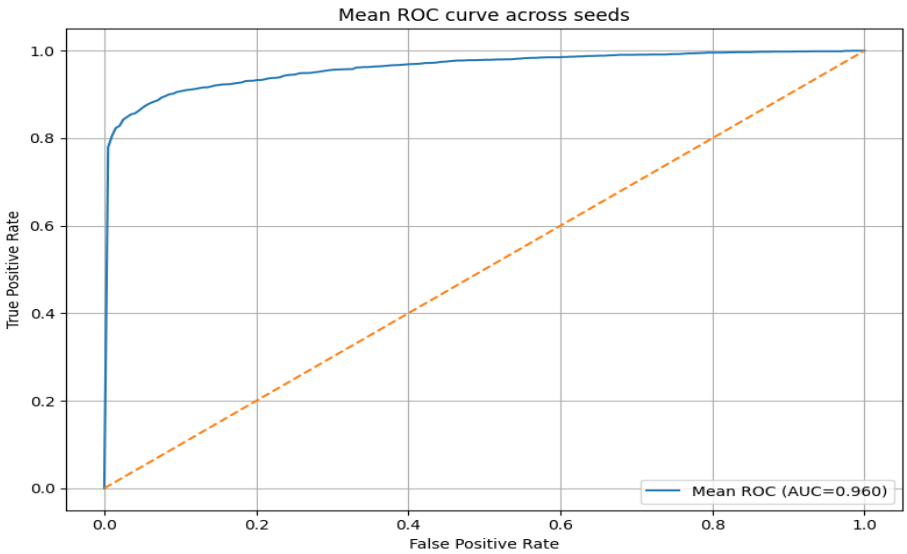
طبق روابط ۲۴ و ۲۵، ROC یک منحنی دوبعدی است که عملکرد تشخیص یک مدل طبقه‌بندی دودویی را با نمایش نرخ مثبت واقعی (TPR)در مقابل نرخ مثبت کاذب (FPR) در سطوح مختلف آستانه تصمیم‌گیری نشان می‌دهد.

معیار AUC-ROC ، که مقداری بین ۰ و ۱ دارد، احتمال این را اندازه‌گیری می‌کند که مدل بتواند به طور تصادفی یک نمونه مثبت را نسبت به یک نمونه منفی رتبه‌بندی کند. مقدار AUC=۱ نشان‌دهنده تشخیص کامل و AUC=۰,۵ معادل یک طبقه‌بندی کننده تصادفی است. در شکل۵ و ۶ منحنی ROC و مقدار AUC برای سید ها و مقدار میانگین نمایش داده شده است.



شکل(۵)، منحنی ROC به ازای هر سید

[۳۷] ROC: Receiver Operating Characteristic
[۳۸] Area under the Receiver Operating Characteristic Curve



شکل(۶)، منحنی ROC میانگین سیدها

۶.نتیجه‌گیری و پیشنهادات

در این پژوهش، یک چارچوب نوین تشخیص ناهنجاری مبتنی بر شبکه خود-رمزگذار گرافی (GAE) برای مقابله با تهدیدات پیچیده‌ای چون حملات تزریق داده جعلی در شبکه‌های توزیع هوشمند ارائه شد. هسته مرکزی این روش، استفاده از قابلیت‌های یادگیری عمیق گرافی برای مدلسازی و یادگیری الگوهای رفتاری شبکه قدرت در شرایط عادی است. مدل پیشنهادی با دریافت داده‌های ولتاژ باس‌ها و ساختار توپولوژیک شبکه، اقدام به یادگیری یک نمایش فشرده و بهینه از حالت سالم سیستم می‌نماید. معیار کلیدی تشخیص در این طرح، محاسبه خطای بازسازی است؛ به گونه‌ای که هرگونه انحراف در داده‌های عملیاتی از الگوی عادی آموخته‌شده، منجر به افزایش محسوس این خطا شده و به عنوان نشانه‌ای از یک حمله یا ناهنجاری شناسایی می‌گردد.

ارزیابی‌های تجربی بر روی شبکه استاندارد IEEE ۳۳-bus و با استفاده از یک مجموعه داده جامع متشکل از ۶۰۰۰۰ سناریو، کارایی این چارچوب را به اثبات رساند. مدل پیشنهادی با دستیابی به دقت بالای ۹۸ درصد و فراخوانی بالای ۹۷ در قالب F۱-Score بالای ۹۸ درصد تبلور یافته است، حاکی از توانایی آن در شناسایی کم‌خطای حملات است. همچنین، نتایج حاصل از ماتریس درهم‌ریختگی به وضوح نشان می‌دهد که مدل به خوبی قادر به تمایز بین داده‌های عادی و داده‌های تحت حمله است و از نرخ هشدار کاذب (False Positive) پایینی برخوردار می‌باشد.

این مدل نیازی به پیش‌دانش دقیق از الگوهای حمله ندارد و صرفاً با یادگیری وضعیت سالم سیستم، قادر به تشخیص هرگونه انحراف مخرب می‌باشد. این ویژگی، چارچوب ارائه‌شده را به یک راهکار امنیتی پیشگیرانه و مقیاس‌پذیر برای پیاده‌سازی در محیط‌های عملیاتی شبکه‌های هوشمند تبدیل می‌کند.

پیشنهادات برای پژوهش‌های آینده:

- افزایش مقاومت در برابر حملات پیشرفته: آموزش مدل با داده‌های آلوده به حملات غیرخطی و چندمرحله‌ای برای افزایش تاب‌آوری در شرایط واقعی.
- استقرار عملیاتی: پیاده‌سازی مدل در محیط‌های شبیه‌سازی بلادرنگ[۳۹] و ارزیابی عملکرد آن در مقیاس بزرگ.
- ترکیب داده ها با نویز: استفاده از نویز جمعی با داده های اصلی برای اموزش مدل برای مقاومت مدل در برابر نویز شبکه.
- استفاده از منابع تولید پراکنده: جایگذاری منابع تولید پراکنده در شبکه اصلی و آزمون مدل در حضور این مناب

منابع

[۱] S. Wang, S. Bi, and Y. J. A. Zhang, "Locational Detection of the False Data Injection Attack in a Smart Grid: A Multilabel Classification Approach," *IEEE Internet Things J*, vol. ۷, no. ۹, pp. ۸۲۱۸–۸۲۲۷, Sep. ۲۰۲۰, doi: ۱۰,۱۱۰۹/JIOT.۲۰۲۰,۲۹۸۳۹۱۱.

[۲] G. Liang, J. Zhao, F. Luo, S. R. Weller, and Z. Y. Dong, "A Review of False Data Injection Attacks Against Modern Power Systems," Jul. ۰۱, ۲۰۱۷, *Institute of Electrical and Electronics Engineers Inc.* doi: ۱۰,۱۱۰۹/TSG.۲۰۱۵,۲۴۹۵۱۳۳.

[۳] S. Bi and Y. J. Zhang, "Using covert topological information for defense against malicious attacks on DC state estimation," *IEEE Journal on Selected Areas in Communications*, vol. ۳۲, no. ۷, pp. ۱۴۷۱–۱۴۸۵, ۲۰۱۴, doi: ۱۰,۱۱۰۹/JSAC.۲۰۱۴,۲۳۳۲۰۵۱.

[٤] Z. Liu, M. Liu, Q. Wang, and Y. Tang, "False Data Injection Attacks on Data-Driven Algorithms in Smart Grids Utilizing Distributed Power Supplies," *Engineering*, vol. ۵۱, pp. ٦۲–۷٤, Aug. ۲۰۲۵, doi: ۱۰,۱۰۱٦/j.eng.۲۰۲٤,۱۱,۰۲۵.

[۵] X. Wang, M. Hu, X. Luo, and X. Guan, "A detection model for false data injection attacks in smart grids based on graph spatial features using temporal convolutional neural networks," *Electric Power Systems Research*, vol. ۲۳۸, Jan. ۲۰۲۵, doi: ۱۰,۱۰۱٦/j.epsr.۲۰۲٤,۱۱۱۱۲٦.

[٦] W. He, W. Liu, C. Wen, and Q. Yang, "Detection of False Data Injection Attacks on Smart Grids Based on A-BiTG Approach," *Electronics (Switzerland)*, vol. ۱۳, no. ۱۰, May ۲۰۲٤, doi: ۱۰,۳۳۹۰/electronics۱۳۱۰۱۹۳۸.

[۳۹] Real-Time

[۷] P. Hu, W. Gao, Y. Li, M. Wu, F. Hua, and L. Qiao, “Detection of False Data Injection Attacks in Smart Grids Based on Expectation Maximization,” *Sensors*, vol. ۲۳, no. ۳, Feb. ۲۰۲۳, doi: ۱۰,۳۳۹۰/s۲۳۰۳۱۶۸۳.

[۸] F. Zhang and Q. Yang, “False data injection attack detection in dynamic power grid: A recurrent neural network-based method,” *Front Energy Res*, vol. ۱۰, Sep. ۲۰۲۲, doi: ۱۰,۳۳۸۹/fenrg.۲۰۲۲,۱۰۰۵۶۶۰.

[۹] Y. Liu, P. Ning, and M. K. Reiter, “False data injection attacks against state estimation in electric power grids,” in *ACM Transactions on Information and System Security*, May ۲۰۱۱. doi: ۱۰,۱۱۴۵/۱۹۵۲۹۸۲,۱۹۵۲۹۹۵.

[۱۰] J. Kim, L. Tong, and R. J. Thomas, “Subspace methods for data attack on state estimation: A data driven approach,” *IEEE Transactions on Signal Processing*, vol. ۶۳, no. ۵, pp. ۱۱۰۲–۱۱۱۴, Mar. ۲۰۱۵, doi: ۱۰,۱۱۰۹/TSP.۲۰۱۴,۲۳۸۵۶۷۰.

[۱۱] H. Haider, A.-M. Zaid Ali, and -Yemen Ayeda Al-Hmadi, “False Data Injection Attacks (FDIA) detection by Deep Learning Techniques in Smart Grids: survey.”

[۱۲] A. Takiddin, R. Atat, M. Ismail, O. Boyaci, K. R. Davis, and E. Serpedin, “Generalized Graph Neural Network-Based Detection of False Data Injection Attacks in Smart Grids,” *IEEE Trans Emerg Top Comput Intell*, vol. ۷, no. ۳, pp. ۶۱۸–۶۳۰, Jun. ۲۰۲۳, doi: ۱۰,۱۱۰۹/TETCI.۲۰۲۲,۳۲۳۲۸۲۱.

[۱۳] T. N. Kipf and M. Welling, “Variational Graph Auto-Encoders,” Nov. ۲۰۱۶, [Online]. Available: http://arxiv.org/abs/۱۶۱۱,۰۷۳۰۸

[۱۴] W. Liao, B. Bak-Jensen, J. Radhakrishna Pillai, Y. Wang, and Y. Wang, “A Review of Graph Neural Networks and Their Applications in Power Systems,” *Journal of Modern Power Systems and Clean Energy*, vol. ۱۰, no. ۲, pp. ۳۴۵–۳۶۰, ۲۰۲۲, doi: ۱۰,۳۵۸۳۳/MPCE.۲۰۲۱,۰۰۰۰۵۸.

[۱۵] A. Howe, D. Peasley, and M. Papa, “Graph Autoencoders for Detecting Anomalous Intrusions in OT Networks Through Dynamic Link Detection,” in *2024 IEEE 21st Consumer Communications & Networking Conference (CCNC)*, IEEE, Jan. ۲۰۲۴, pp. ۱–۶. doi: ۱۰,۱۱۰۹/CCNC۵۱۶۶۴,۲۰۲۴,۱۰۴۵۴۸۴۱.

[۱۶] C. Zhang and J. W. Jung, “Enhanced Graph Autoencoder for Graph Anomaly Detection Using Subgraph Information,” *Applied Sciences (Switzerland)*, vol. ۱۵, no. ۱۵, Aug. ۲۰۲۵, doi: ۱۰,۳۳۹۰/app۱۵۱۵۸۶۹۱.

[۱۷] R. Chalapathy and S. Chawla, “Deep Learning for Anomaly Detection: A Survey,” Jan. ۲۰۱۹, [Online]. Available: http://arxiv.org/abs/۱۹۰۱,۰۳۴۰۷

«نشریه علمی (فصلنامه) «انرژی ایران
دوره ۲۸، شماره۱، بهار ۱۴۰۴، صفحه ۶۵-۸۱

A Deep Learning-Based Approach for Detecting False Data Injection Attacks in Smart Grids

Authors:
Alireza Soltani[1], Hamidreza Baghaee-Kashi[2], Mahmoudreza Haghifam[3]

[1] M.Sc. Student, Electrical Engineering Department, Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran.
Email: alireza_soltani@modares.ac.ir

[2] Assistant Professor, Electrical Engineering Department, Faculty of Electrical and Computer Engineering, Tarbiat Modares University, Tehran, Iran.

[3] Professor, Electrical Engineering Department, Faculty of Electrical and Computer Engineering, Tarbiat Modares, University,Tehran, Iran.

Abstract

The proliferation of smart distribution networks and the escalation of cyber threats have rendered the security of power systems critically imperative. Among the most disruptive threats, False Data Injection (FDI) attacks compromise grid stability by evading conventional detection mechanisms. This research proposes a **Graph Autoencoder-based detection model** that proactively identifies such attacks in the IEEE ۳۳-bus distribution system. Designed within a deep learning architecture, the model harnesses the representational capabilities of Graph Neural Networks (GNNs) to learn normal operational patterns of the grid and subsequently detect anomalies through reconstruction error computation. The results demonstrate that the proposed model effectively identifies intrusions with **over ۹۸٪ accuracy**, thereby substantially enhancing the performance of the protection system.

Submission Date:
۲۰۲۵/۱۱/۰۳
Accepted Date:
۲۰۲۶/۰۲/۲۵

Keywords:
Smart grid
Anomaly Detection
Cyber-Attack
Deep Learning