فصلنامه تأمین اجتماعی ۷۳
سال بیستم ● شماره سوم ● پاییز ۱۴۰۳

ارائه مدلی برای پیش‌بینی فرایندهای سازمان تأمین اجتماعی با استفاده از یادگیری عمیق بازگشتی

مهرداد فدائی پله شاهی[۱]، امیرمحمد فدائی پله شاهی[۲]

چکیده

هدف: دستاوردهای اخیر در هوش مصنوعی، فرصت‌های جدیدی را در بیمه‌های اجتماعی ایجاد کرده تا خدمات متناسب با تکیه بر دانش جدید به بیمه‌شدگان ارائه گردد. بااین‌حال، صنعت بیمه‌های اجتماعی برای استفاده از هوش مصنوعی با چالش‌های مهمی از جمله داده‌های ناهمگن، توزیع نامتوازن داده‌ها در دسته‌های مورد پیش‌بینی، نرخ پایین تخصیص هر داده به یک دسته و وجود ویژگی‌های فراوان در یک فرایند روبرو است. هدف مطالعه حاضر ارائه روشی کارآمد برای پیش‌بینی فرایندهای بیمه‌های اجتماعی است تا باتوجه‌به چالش‌های داده‌های ناهمگن، نامتوازن و با ویژگی‌های فراوان، بتواند نتایج دقیق و بهینه‌تری را در پیش‌بینی نتایج فرایندهای بیمه‌ای ارائه دهد

روش: در این پژوهش، روشی نوین مبتنی بر شبکه‌های یادگیری عمیق بازگشتی با ساختار حافظه بلند -کوتاه‌مدت (LSTM) و بهره‌گیری از مراحل پیش‌پردازش داده و تقسیم‌بندی داده‌ها به دسته‌های واحد A، برای مقابله با چالش‌های موجود ارائه شده است. هدف اصلی این روش، بهبود دقت پیش‌بینی فرایندهای بیمه‌های اجتماعی با هزینه محاسباتی کمتر و با تکیه بر سوابق گذشته فرایندها است.

یافته‌ها: روش پیشنهادی با داده‌های حقیقی سازمان تأمین اجتماعی ایران شبیه‌سازی‌شده و نتایج نشان می‌دهد که استفاده از روش ارائه شده، میزان استفاده از منابع مصرفی شامل حافظه و مقدار استفاده از واحد پردازشگر مرکزی را به میزان بسیار جزئی افزایش داده ولی میزان خطا، نسبت به دو روش مورد مقایسه، با کاهش چشمگیری مواجه شده است. همچنین کاهش خطا حتی با داده‌های کمتر در مرحله آموزش یادگیری عمیق، از دستاوردهای روش پیشنهادی نیز است.

نتیجه‌گیری: روش ارائه شده، توانسته است با حفظ صرفه‌جویی نسبی در منابع پردازشی، دقت پیش‌بینی را به شکل چشمگیری افزایش دهد و کاهش خطا را تسریع کند؛ بنابراین این مدل برای پیش‌بینی فرایندهای بیمه‌های اجتماعی، گزینه‌ای مناسب و اثربخش به‌شمار می‌رود و قابلیت

۱- دکتری علوم کامپیوتر گرایش محاسبات نرم و هوش مصنوعی، کارشناس معاونت بیمه‌ای سازمان تأمین اجتماعی. (نویسنده مسئول)
m.fadaei.1357@gmail.com

۲- دانشجوی رشته علوم کامپیوتر، دانشکده علوم و فناوری‌های همگرا و کوانتوم، دانشگاه آزاد اسلامی واحد تهران مرکزی، تهران، ایران

استفاده در سازمان‌های بیمه‌ای برای هوشمندسازی و بهبود عملکرد را دارد.

واژه‌های کلیدی: حافظه بلند -کوتاه‌مدت، یادگیری عمیق بازگشتی، پیش‌بینی، تأمین اجتماعی، شبکه عصبی.

مقدمه

پیش‌بینی و برآورد در کسب‌وکار سازمان‌ها از اهمیت بالایی برخوردار است و نتایج پیش‌بینی‌های دقیق منجربه افزایش بهره‌وری، صرفه‌جویی در هزینه‌ها، بهبود کمیت و کیفیت سود و همچنین ارائه خدمات بهتر به ذی‌نفعان می‌شود.

یکی از محورهای اصلی در پژوهش‌های اخیر، بررسی اثربخشی یادگیری عمیق در مدل‌سازی داده‌های رفتاری مشتریان به‌صورت فردی است. به‌ویژه در حوزه امور مالی خرد که حوزه‌ای کلیدی در تحقیقات عملیاتی محسوب می‌شود، پیش‌بینی رفتارهای پرخطر معامله‌گران با دقت بیشتری توسط مدل‌های یادگیری عمیق انجام شده است. نتایج تجربی نشان می‌دهد که این روش‌ها نسبت به الگوریتم‌های سنتی یادگیری ماشین عملکرد بهتری دارند (یانگ[1] و همکاران، ۲۰۱۸).

با افزایش پیچیدگی و حجم داده‌های کسب‌وکاری، پیش‌بینی فرایندهای کسب‌وکار به عنوان یکی از موضوعات مهم در حوزه مدیریت فرایندها و تحلیل داده‌ها مطرح شده است. هدف اصلی پیش‌بینی فرایندهای کسب‌وکار، تحلیل داده‌های تاریخی مرتبط با فرایندها برای شناسایی الگوها و روندهای آینده است که می‌تواند به بهبود تصمیم‌گیری و افزایش کارایی سازمان‌ها کمک کند.

باتوجه‌به اهمیت پیش‌بینی دقیق فرایندهای کسب‌وکار و توانمندی‌های شبکه‌های عصبی بازگشتی در پردازش داده‌های توالی‌دار، ارائه مدلی مبتنی بر یادگیری عمیق بازگشتی می‌تواند گام مهمی در بهبود عملکرد سیستم‌های مدیریت فرایندهای کسب‌وکار باشد.

هدف اصلی این تحقیق، ارائه مدلی مبتنی بر یادگیری عمیق بازگشتی برای پیش‌بینی فرایندهای کسب‌وکار است.

مدل‌های پیش‌بینی ورشکستگی به طور معمول با استفاده از نسبت‌های متغیرهای حسابداری مانند نسبت کل دارایی به کل بدهی، احتمال پیش‌فرض را برآورد می‌کنند (گنگ[2] و همکاران، ۲۰۱۵). در چهارچوب یادگیری عمیق، این نسبت‌ها به عنوان بازنمایی‌های سطح پایین داده‌ها تلقی می‌شوند. لایه‌های ابتدایی شبکه‌های عصبی عمیق قادرند ارتباط میان متغیرها را به صورت خودکار یاد گرفته و نسبت‌های مالی را به شکل داده‌محور استخراج کنند. در سطوح بالاتر، این شبکه‌ها می‌توانند الگوهای پیچیده‌تر و انتزاعی مانند وابستگی‌های متقابل متغیرها یا مفهوم «وام‌گیرنده بزهکار» را مدل‌سازی کنند. این توانایی یادگیری بازنمایی‌های سلسله‌مراتبی، به شبکه‌های عمیق امکان می‌دهد الگوهایی را کشف کنند که در داده‌های آموزشی کمتر دیده شده‌اند، چالشی که در مدل‌های سنتی یادگیری

1- Yang
2- Geng

ماشین رایج است (بنگیو[1]، ۲۰۰۹).

موفقیت چشمگیر یادگیری عمیق در زمینه‌هایی همچون بینایی ماشین، پردازش زبان طبیعی و سایر حوزه‌ها (اسچمیدبولر[2]، ۲۰۱۵)، منجر به توسعه کاربردهایی شده که به داده‌های غیرساختاری وابسته‌اند (لیو[3] و همکاران، ۲۰۱۷).

در صنعت بیمه نیز تحولات اخیر در حوزه هوش مصنوعی فرصت‌های جدیدی را برای ارائه خدمات دانش‌بنیان، بهبود عملیات و عملکردهای تجاری فراهم کرده است. با این حال، داده‌های بیمه‌های تجاری و اجتماعی معمولاً ناهمگن، نامتوازن و با ابعاد بالا هستند که چالش‌های قابل توجهی در استفاده از روش‌های یادگیری ایجاد می‌کنند. الگوریتم‌های یادگیری ماشین سنتی اغلب برای داده‌های همگن و متعادل طراحی شده‌اند، درحالی‌که افزایش قدرت محاسباتی و پیشرفت در الگوریتم‌های یادگیری عمیق امکان تحلیل داده‌های پیچیده‌تر و متنوع‌تر را فراهم آورده است.

برخی بیمه‌گذاران با بهره‌گیری از روش‌های یادگیری ماشین قادرند داده‌های بزرگ را با هزینه کمتر تحلیل کرده و سودآوری کسب‌وکار خود را بهبود بخشند. به عنوان نمونه، این روش‌ها می‌توانند در فرایند تعهدگیری، کمک به کارکنان برای مرتب‌سازی داده‌های بزرگ، شناسایی موارد پرخطر و کاهش ادعاهای بیمه‌ای مؤثر باشند (جیانگ[4] و همکاران، ۲۰۱۸).

با این وجود، استفاده از تکنیک‌های هوش مصنوعی در داده‌های متنوع بیمه‌های اجتماعی همچنان برای بسیاری از سازمان‌ها دشوار است. چهار چالش کلیدی در این زمینه عبارت‌اند از:

- وجود ویژگی‌های فراوان در یک پدیده که ثبت و استفاده از آن را به عنوان یک مجموعه نظام‌مند در یک ساختار هوش مصنوعی با مشکل مواجه می‌کند.

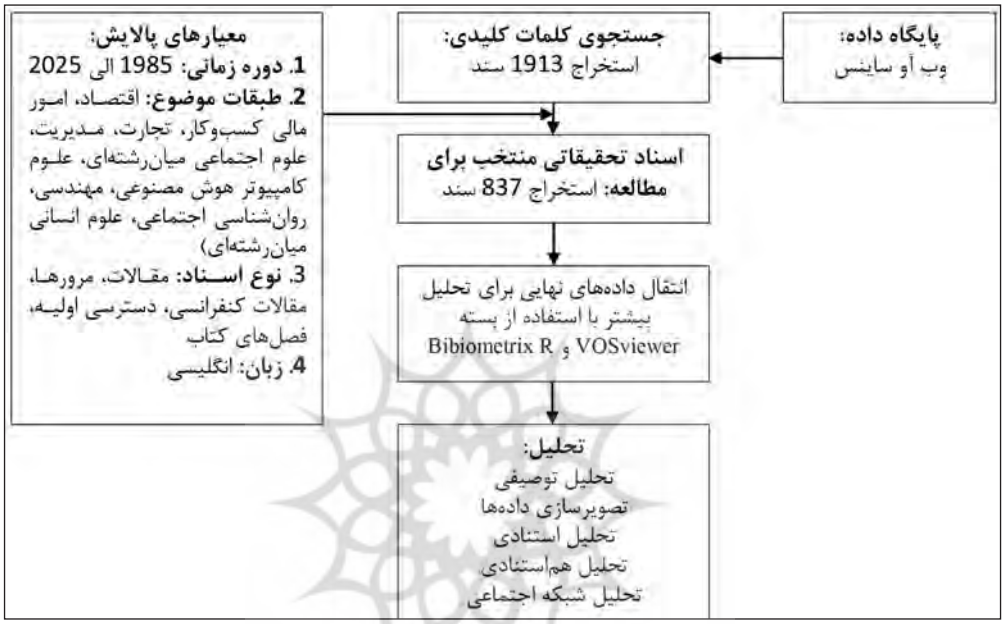
سازمان‌های بیمه‌های اجتماعی، اطلاعات بسیار گسترده‌ای از بیمه‌شدگان و خدمات قابل ارائه به منظور کنترل ریسک و ارائه خدمات مطلوب‌تر، جمع‌آوری می‌کنند. در مجموعهٔ داده‌های مذکور، ابعاد معمول و مشخصات بیمه‌شده و موارد مرتبط با خدمات می‌تواند بیش از ده‌ها ویژگی، ارائه شود.

در مجموعهٔ داده‌های بیمه‌های اجتماعی، کیفیت داده و تراکم داده‌ها می‌تواند کاملاً مشهود باشد. انتخاب مؤثرترین ویژگی‌های یک هدف کسب‌وکار بیمه‌های اجتماعی همراه با ارتباط‌های مؤثر برای پردازش بهینه داده‌ها، امری بسیار چالش‌برانگیز است. به‌عنوان‌مثال، نسبت سن و سابقه ممکن است عامل مهمی در تجزیه و تحلیل ریسک تعهدات باشد، درحالی‌که میزان حق‌بیمه پرداختی یکی از مهم‌ترین ویژگی‌ها

1- Bengio
2- Schmidhuber
3- Liu
4- Jiang

در آنالیز ریسک‌های قابل محاسبه دخیل می‌باشد. استفاده مؤثر از محدود ویژگی‌هایی که بیشترین ارتباط را در میان ویژگی‌های با ابعاد بالا داشته باشند، عامل بسیار مهمی در تجزیه و تحلیل داده‌های بیمه‌های اجتماعی محسوب می‌شود.

ناهمگنی داده‌ها



شکل ۱. مثالی از مجموعه داده‌های ناهمگن در بیمه‌های اجتماعی

داده‌های ناهمگن بیمه‌های اجتماعی به دو سطح اصلی تقسیم می‌شود: سطح داده و سطح ساختار (پاولیدیس[1] و همکاران، ۲۰۰۱) (هو[2] و همکاران، ۲۰۱۷) (ایمرت[3] و همکاران، ۲۰۱۴) و ناهمگنی سطح داده شامل انواع مختلفی از داده‌ها (مانند ساختار داده که حاوی انواع مختلفی از داده؛ مانند عدد صحیح و کاراکتر است) بوده، درحالی‌که ناهمگنی سطح ساختار ترکیبی از انواع مختلف داده و منابع داده می‌باشد (به‌عنوان مثال، ساختار پیچیده داده که ترکیبی از ویژگی‌های ایستا برای تشریح ویژگی بوده و خدمات پویا برای فعالیت‌های سری زمانی است). داده‌های ناهمگن، معمولاً جایگاه‌های مختلفی را در فضای داده، اشغال می‌کنند؛ بنابراین با استفاده از یک روش یادگیری عمیق یا یک پیش‌بینی به

1- Pavlidis
2- Hu
3- Emmert

منظور استخراج الگوها از داده‌های ناهمگن، مقایسه مفید ارائه نمی‌شود. (میکولو[1] و همکاران، ۲۰۱۳) (رودولف[2] و همکاران، ۲۰۱۶) (گریناکر[3]، ۱۹۸۴) (گلوبرسون[4] و همکاران، ۲۰۰۶) نحوه استفاده مؤثر از این اطلاعات ناهمگن در بیمه‌های اجتماعی، یک سؤال بسیار مهم و کاربردی می‌باشد.

نرخ پایین تخصیص هر داده به یک دسته

در مقایسه با دیگر فرایندهای تجاری مالی؛ مانند بانکداری که در آن تراکنش حساب‌های بانکی مشتریان، معمولاً با تعدد بالا انجام می‌شود و با ایجاد توالی طولانی، رفتارهای مشتری ردیابی می‌شود، فرکانس درخواست‌ها و ارائه خدمات به بیمه‌شدگان در بیمه‌های اجتماعی بسیار پایین‌تر است. در بیمه‌های اجتماعی، درخواست و یا حضور فیزیکی یا امور غیرحضوری مستقیم ممکن است تنها یک یا دو بار در سال صورت گرفته و ارائه خدمات در یک چرخه عمر مانند درخواست‌های بیمه‌شدگان نیز با نرخ پایین رخ می‌دهد. هر چند که ممکن است این واقعیت نادیده گرفته شود که پیگیری درخواست یا ارائه خدمات برای عملیات‌های بیمه‌های اجتماعی مهم بوده و تماس بیمه‌شده با سازمان‌های بیمه‌های اجتماعی نشان‌دهنده یک رفتار مهم مانند پرداخت حق‌بیمه، درخواست خدمات و غیره بوده و بنابراین، تجزیه و تحلیل عمیق داده‌های مربوط به تراکنش‌های بیمه‌شدگان، بینش دقیق‌تری از بیمه‌های اجتماعی ارائه می‌کند.

توزیع نامتوازن داده‌ها در دسته‌های مورد پیش بینی

عملیات بیمه‌های اجتماعی بر روی طبقات اقلیت به جای طبقه‌بندی متعادل با فناوری‌های سنتی متمرکز است. ریسک‌های قابل تعریف در بیمه‌های اجتماعی، معمولاً رویدادهایی بوده که به احتمال کمی رخ می‌دهد که این امر بدین معنی است که اهداف کسب‌وکار آنالیز شده معمولاً بر روی طبقه‌بندی اقلیت در بیمه‌های اجتماعی متمرکز است. در مقایسه با طبقه‌بندی متعادل سنتی، توزیع نامتوازن داده‌ها در دسته‌های مورد پیش بینی را خواهیم داشت (هی[5] و همکاران، ۲۰۰۹) (پرووست[6]، ۲۰۰۰).

به منظور استفاده از فرصت‌های هوش مصنوعی، بیمه‌های اجتماعی نیازمند یک چهارچوب یادگیری ماشین مؤثر به منظور غلبه بر این چالش‌ها مانند پیش بینی بوده و روابط معنی‌داری در عملیات بیمه‌های اجتماعی ایجاد می‌کنند. در مقایسه با یادگیری ماشین سنتی، رویکردهای یادگیری

1- Mikolov
2- Rudolph
3- Greenacre
4- Globerson
5- He
6- Provost

عمیق در پیش بینی دارای مزایایی بوده و قادر به استخراج ویژگی‌ها و روابط غیرخطی بدون تکیه بر فرضیه‌های اقتصاد سنجی و تخصص انسانی است (وانگ[1] و همکاران، ۲۰۱۶) (یانگ و همکاران، ۲۰۱۶) (پان[2] و همکاران، ۲۰۱۸) (وانگ و همکاران، ۲۰۱۷) (پان و همکاران، ۲۰۱۶).

۲- مفاهیم اولیه

بیمه‌های اجتماعی، صندوق‌های بازنشستگی و همچنین بیمه‌های تجاری، تفاوت‌های بسیاری از نظر جزئیات و سازوکار و مدل‌های کسب‌وکار دارند؛ ولی همگی آنها برای مقابله با بحران‌های پیش رو شکل گرفته و دارای اهداف نهایی مشترکی هستند؛ بنابراین، پیش‌بینی صحیح در کسب‌وکارهای بیمه‌ای و علی‌الخصوص بیمه‌های اجتماعی، دارای اهمیت فراوانی بوده و در ماندگاری مدل کسب‌وکار، نقش بسیار زیادی خواهند داشت. در این بخش، چند مفهوم مفید در این حوزه توضیح داده می‌شود

۲-۱. یادگیری عمیق

مفهوم «عمق» در بسیاری از مطالعات تجربی، به‌ویژه در زمینه‌هایی مانند شبکه‌های عصبی مصنوعی و ماشین‌های بردار پشتیبان، به عنوان عاملی کلیدی در بهبود عملکرد مدل‌ها نسبت به رگرسیون‌های ساده مطرح شده است (لسمان[3] و همکاران، ۲۰۱۵). افزایش عمق در این مدل‌ها، امکان یادگیری سطوح بالاتری از بازنمایی داده را به‌صورت ضمنی فراهم می‌سازد. این سطوح بازنمایی پیچیده‌تر، موجب تعمیم بهتر مدل به ترکیب‌های جدیدی از ویژگی‌ها می‌شود که ممکن است در داده‌های آموزشی کمتر مشاهده شده باشند. به‌علاوه، ظرفیت بالاتر مدل‌های عمیق اجازه می‌دهد تا مدل تغییرات بیشتری را در تابع هدف فراگیرد؛ موضوعی که منجر به تمایز دقیق‌تر بین کلاس‌ها می‌گردد.

از سوی دیگر، تعداد واحدهای محاسباتی که یک مدل می‌تواند از آن بهره‌مند شود، به طور مستقیم تحت‌تأثیر اندازه مجموعه داده‌های آموزشی است. در معماری‌های کم‌عمق، برای مدل‌سازی دقیق تابع هدف، نیاز به تعداد بسیار بیشتری از واحدهای محاسباتی وجود دارد؛ عاملی که به نوبه خود، مستلزم بهره‌گیری از داده‌های آموزشی گسترده‌تر نسبت به مدل‌های عمیق است (بنگیو[4]، ۲۰۰۹).

1- Wang
2- Pan
3- Lessmann
4- Bengio

۲-۲. شبکه عصبی بازگشتی (RNN[1]) و حافظه کوتاه-بلندمدت (LSTM[2])

حافظه کوتاه‌مدت بلند (LSTM)، گونه‌ای خاص از شبکه‌های عصبی بازگشتی (RNN) است که برای یادگیری وابستگی‌های بلندمدت در داده‌های توالی‌دار طراحی شده‌اند و در مسائل پیش‌بینی عملکرد قابل توجهی دارند (هوچریتر[3] و همکاران، ۱۹۹۷). یکی از مزایای کلیدی حافظه کوتاه‌مدت بلند (LSTM)، مقاومت در برابر مسئله محوشدگی گرادیان است؛ مشکلی که در شبکه‌های عصبی بازگشتی سنتی هنگام یادگیری وابستگی‌های طولانی‌مدت به طور جدی رخ می‌دهد. ساختار حافظه کوتاه‌مدت بلند به‌طور خاص برای مقابله با این چالش طراحی شده است (جوزف ویز[4] و همکاران، ۲۰۱۵).

درحالی‌که ساختار سلولی ساده در شبکه‌های عصبی بازگشتی برای نگهداری اطلاعات در طول زمان ناکارآمد است، معماری حافظه کوتاه‌مدت بلند با استفاده از حافظه سلولی و دروازه‌های کنترل‌گر امکان ذخیره و بازیابی اطلاعات در بازه‌های زمانی طولانی‌تر را فراهم می‌سازد. یک سلول LSTM مبنا به شرح زیر تعریف می‌شود: پذیرش C_{t-1} و h_{t-1} به عنوان حالت و اطلاعات ورودی از سلول بازشده اولیه در همان سطح و پذیرش X_t به عنوان ورودی از سلول‌های لایه‌های قبلی می‌باشد. به نوبه خود، C_t و h_t به عنوان حالت و اطلاعات خروجی به سلول باز شده توالی منتقل می‌کند و h_t را به عنوان خروجی به لایه بعدی ارائه می‌دهد (اورمان[5] و همکاران، ۲۰۱۶) .

۳- چهارچوب نظری پژوهش

در سال‌های اخیر، با افزایش پیچیدگی و حجم داده‌های کسب‌وکاری، پیش‌بینی فرایندهای کسب‌وکار به عنوان یکی از موضوعات مهم در حوزه مدیریت فرایندها و تحلیل داده‌ها مطرح شده است. هدف اصلی پیش‌بینی فرایندهای کسب‌وکار، تحلیل داده‌های تاریخی مرتبط با فرایندها برای شناسایی الگوها و روندهای آینده است که می‌تواند به بهبود تصمیم‌گیری و افزایش کارایی سازمان‌ها کمک کند.

پژوهش‌های متعددی در زمینه پیش‌بینی فرایندهای کسب‌وکار با استفاده از روش‌های یادگیری عمیق انجام شده است. به طور ویژه، شبکه‌های عصبی بازگشتی به دلیل توانایی‌شان در مدل‌سازی داده‌های توالی‌دار، مورد توجه قرار گرفته‌اند.

1- Recurrent Neural Networks
2- Long Short Term Memory
3- Hochreiter
4- Jozefowicz
5- Evermann

در سال‌های اخیر، استفاده از مدل‌های پیش‌بینی مبتنی بر احتمالات به‌عنوان ابزاری مؤثر برای ردیابی پیشرفت فرایندهای پیچیده، مورد توجه گسترده قرار گرفته است. این مدل‌ها قادرند احتمال وقوع یک فعالیت خاص را باتوجه‌به وضعیت جاری و اطلاعات مربوط به سایر زیرفرایندها برآورد کنند. برای نمونه، پففر[1] (۲۰۰۵) مدلی احتمالی برای نمایش فرایند ارائه داده است که در آن از شبکه‌های بیزین پویا (Dynamic Bayesian Networks) جهت استنتاج و پیش‌بینی در فرایندهای حاوی ساختارهای تکراری و موازی استفاده می‌شود.

لکشمانان و همکاران[2]، (۲۰۱۵)، با استخراج مدل مارکوفی از فرایندهای نیمه‌ساخت‌یافته مبتنی بر داده‌های واقعی، مدلی توسعه دادند که قادر به پیش‌بینی فعالیت‌های آینده در فرایند کسب‌وکار بر اساس داده‌های همراه هر رویداد است؛ این مدل به‌ویژه اطلاعات مربوط به حلقه‌ها و ساختارهای موازی را نیز در تحلیل خود لحاظ می‌کند.

اورمان و همکاران، (۲۰۱۶) با بهره‌گیری از تکنیک‌های یادگیری عمیق و به‌طور خاص، شبکه‌های عصبی بازگشتی (RNN) با سلول‌های LSTM، مدلی برای پیش‌بینی رویداد بعدی در یک فرآیند در حال اجرا توسعه داده‌اند. این رویکرد با الهام از پردازش زبان طبیعی توانسته است رفتارهای دینامیک فرآیند را با دقت بیشتری مدل‌سازی کند.

در مطالعات متعددی نیز از داده‌های ثبت‌شده توسط سیستم‌های اطلاعاتی جهت آموزش مدل‌های پیش‌بینی استفاده شده است. مارکوس و همکاران[3] (۲۰۱۷) با تحلیل داده‌های لاگ رویدادها، چهارچوبی برای نظارت و پیش‌بینی در فرآیندهای کسب‌وکار ارائه کردند که منجر به بهبود مدیریت بحران و تصمیم‌گیری شده است. این مطالعات از الگوریتم‌های متنوعی شامل رگرسیون، درخت تصمیم، شبکه‌های عصبی و مدل‌های احتمالاتی بهره برده‌اند و بیش از ۳۹ اثر پژوهشی را در این حوزه مرور کرده‌اند.

در زمینه تحلیل شبکه‌های اجتماعی نیز مدل‌های گرافیکی احتمالاتی (PGMs[4]) ـ شامل ساختارهای جهت‌دار و بدون جهت توسط فراست[5] و همکاران (۲۰۱۵) پیشنهاد شده‌اند که از انعطاف‌پذیری و دقت پیش‌بینی بالایی برخوردارند.

1- Pfeffer
2- Lakshmanan et all
3- Márquez
4- Probabilistic Graphical Models

در حوزه‌هایی مانند مدل‌سازی زبان طبیعی (کالبرت[1] و همکاران، ۲۰۰۸)، یا خلق اثر موسیقی (بریوت و همکاران، ۲۰۱۷)، توصیه‌گرهای موسیقی (وانگ[2] و همکاران، ۲۰۱۴)، پیش‌بینی فعالیت‌های محاسبات ابری (کومار[3] و همکاران، ۲۰۱۸)، و پیش‌بینی رفتار مشتریان در شبکه‌های اجتماعی (نیمی[4] و همکاران، ۲۰۱۷) نیز از شبکه‌های عصبی بازگشتی و حافظه کوتاه‌مدت بلند به‌صورت گسترده استفاده شده است. علاوه‌برآن، الگوریتم‌های استخراج قواعد همبستگی نیز برای کاربردهای گوناگون، نظیر تحلیل داده‌های ژنی (جیانگ[5] و همکاران، ۲۰۰۵) و داده‌کاوی در وبلاگ‌ها (هوانگ[6] و همکاران، ۲۰۰۲)، نتایج موفقی را نشان داده‌اند. گیبسون[7] و همکاران (۲۰۱۴) نیز شبکه‌ای عصبی مبتنی بر رویداد با تأخیرهای قابل یادگیری را برای پیش‌بینی دنباله‌های زمانی معرفی کرده‌اند.

بسیاری از رویکردهای پیش‌بینی به جای تمرکز بر پیش‌بینی رویداد بعدی، بر تخمین زمان باقی‌مانده تا تکمیل فرایند متمرکز شده‌اند. دانگن[8] و همکاران، (۲۰۰۸) از رگرسیون افزایشی در تکرارهای رویداد و زمان‌های اجرایی برای این منظور بهره بردند. همچنین، پاندی[9] و همکاران، (۲۰۱۱) با مدل‌سازی توالی رویدادها و زمان اجرا از طریق مدل پنهان مارکوف، به پیش‌بینی رفتار فرایند پرداختند.

لین[10] و همکاران (۲۰۱۹) یک مدل تلفیقی بر پایه RNN معرفی کرده‌اند که اطلاعات رویداد و ویژگی‌های آن را به‌صورت جداگانه با LSTM کدگذاری کرده و سپس با استفاده از یک لایه رمزگشا، رویداد بعدی و ویژگی‌های آن را به‌طور هم‌زمان پیش‌بینی می‌کند.

همچنین جیانگ[11] و همکاران (۲۰۱۸) چهارچوبی موازی و مقرون‌به صرفه با هدف بهینه‌سازی عملیات بیمه‌ای ارائه داده‌اند که نیازی به پیش‌پردازش داده ندارد و از شبکه‌های عصبی موازی برای مدل‌سازی طبقه‌بندی استفاده می‌کند.

با رشد سریع تحقیقات در حوزه یادگیری عمیق، کاربردهای این فناوری در تحلیل داده‌های مالی، بیمه‌ای و شبکه‌های اجتماعی نیز گسترش یافته است (اسچمیدبوبر[12]، ۲۰۱۵)، (لین و همکاران،

1- Collobert
2- Wang
3- Kumar
4- Niimi
5- Jiang
6- Huang
7- Gibson
8- Dongen
9- Pandey
10- Lin
11- Jiang
12- Schmidhuber

۲۰۱۷)، (لکان[1] و همکاران، ۲۰۱۵). هیتون[2] و همکاران، (۲۰۱۷) از شبکه‌های عصبی برای پیش‌بینی ریسک و مدل‌سازی پویایی در داده‌های دفتر سفارش استفاده کرده‌اند. در مطالعات دیگر، از یادگیری عمیق برای پیش‌بینی‌های مالی مبتنی بر داده‌های متنی بهره گرفته شده است (کروس و همکاران، ۲۰۱۷).

سریگنانو[3] و همکاران (۲۰۱۶) با استفاده از شبکه‌های عصبی بازگشتی تا ۷ لایه، انتقال وام‌های فردی به وضعیت‌های مختلف مانند بازپرداخت، معوقه یا کلاهبرداری را مدل‌سازی کرده‌اند. معماری استفاده‌شده در این پژوهش، شامل لایه‌های متنوعی از واحدهای عصبی و استفاده از پیش آموزش بدون نظارت برای استخراج ویژگی‌های پیش‌بینی بوده است که در مطالعات مالی اثربخشی قابل توجهی داشته‌اند.

۴- روش‌شناسی پژوهش

در این پژوهش، با هدف ارائه مدلی کارآمد برای پیش‌بینی فرایندهای کسب‌وکار در حوزه بیمه‌های اجتماعی، یک رویکرد نوین مبتنی بر شبکه‌های یادگیری عمیق بازگشتی با ساختار حافظه بلند-کوتاه‌مدت (LSTM) طراحی و پیاده‌سازی شده است. این مدل با بهره‌گیری از مراحل پیش‌پردازش داده‌ها، نرمال‌سازی ویژگی‌ها و تقسیم‌بندی توالی‌های زمانی به ورودی‌های قابل‌پردازش برای شبکه عصبی، توسعه یافته است.

برای بررسی مدل پیشنهادی و آزمون فرضیه‌های پژوهش، از یک مجموعه‌داده واقعی و نامتوازن متعلق به سازمان تأمین اجتماعی ایران در حوزه درخواست ازکارافتادگی و تشخیص کمیسیون پزشکی استفاده شده است. این داده‌ها شامل ۸۴۶ نمونه به‌صورت تصادفی از سال ۲۰۱۹ و از معاونت درمان سازمان تأمین اجتماعی جمع‌آوری شده‌اند. داده‌های مورد استفاده شامل ترکیبی از ویژگی‌های رشته‌ای (categorical) و عددی (numerical) هستند.

روش پیشنهادی برای ساخت مدل پیش بینی مبتنی بر یادگیری عمیق بازگشتی با مراحل محدود و با پیش‌پردازش ارائه شده و تقسیم داده‌ها به اندازه‌های A دسته واحد برای تغییر یک حافظه بلند-کوتاه‌مدت (LSTM) و حل چالش‌های موجود پیشنهاد گردیده است. صنعت بیمه‌های اجتماعی برای استفاده از هوش مصنوعی با ۴ چالش عمده روبرو می‌باشد:

- داده‌های ناهمگن
- توزیع نامتوازن داده‌ها در دسته‌های مورد پیش‌بینی

1- LeCun
2- Heaton
3- Sirignano

- نرخ پایین تخصیص هر داده به یک دسته
- وجود ویژگی‌های فراوان در یک پدیده که ثبت و استفاده از آن را به عنوان یک مجموعه نظام‌مند در یک ساختار هوش مصنوعی با مشکل مواجه میکند.

چالش‌های داده‌های بیمه‌ای باتوجه‌به اینکه این نوع داده‌ها از قابلیت ترمیمی کمی برخوردارند به این معنی که برای مثال فقط ممکن است که یک‌بار در سال از مشتریان یا صاحبان کسب‌وکار گرفته شود، بیشتر نمایان خواهد شد.

الگوریتم‌های یادگیری ماشین با رویکرد سنتی معمولاً فقط در مجموعه داده‌های استاندارد، عملکرد قابل توجیهی دارند و برای داده‌های معمولاً یک‌دست و متعادل مناسب‌تر هستند. بهره‌گیری از یادگیری عمیق تا حدودی این روند را باتوجه‌به پتانسیل موجود در این روش‌ها و برخی از این چالش‌ها، تسهیل می‌نماید. اما طراحی یک شیوه حل مسئله با شبکه یادگیری عمیق، خود نیازمند دقت در منابع مصرفی و میزان کارآمدی استفاده از این منابع در مقابل نتایج حاصل شده می‌باشد. در این مقاله روشی مبتنی بر یادگیری عمیق بازگشتی با مراحل محدود و با پیشپردازش ارائه می‌شود.

شبکه عصبی بازگشتی پیشنهادی برای ساخت مدل پیش بینی، نوعی از شبکه‌های چند به چند بر خط است. در این نوع از شبکه چند لایه، هر سطح به صورت برخط هم با گرههای لایه L و هم با گرههای لایه L+۱ در ارتباط است. این شبکه هم‌زمان با دریافت هر ورودی، پردازش می‌کند و خروجی جدید تحویل خواهد داد و همچون تمامی شبکه‌های عصبی بازگشتی دارای یک وضعیت داخلی است و در هر مرحله برای تصمیمگیری، علاوه بر ورودی جدید به آن وضعیت هم رجوع می‌کند. که با هر ورودی جدید به‌روزرسانی خواهد شد.

برای حل چالش‌های موجود نیازمندیم که در ساختار یک شبکه بازگشتی، تغییراتی به وجود آوریم. از آنجایی که داده‌های ورودی دارای عدم توازن در ساختار خود هستند باید قبل از آنکه داده‌ها وارد روند آموزش گردند، یک مرحله پیش آموزش روی داده‌ها انجام گیرد. این رخداد (پیش آموزش) در هر LSTM، روی داده‌ها صورت میگیرد. این مرحله به دو دلیل به LSTM اضافه می‌گردد:

- افزایش نرخ تفکیک به‌منظور افزایش میزان تفاوت نرخ تخصیصی هر داده به یک دسته
- کاهش میزان منابع مصرفی با کاهش انجام محاسبات

اگر بتوان مدلی یافت که در آن هر بازه از داده‌های ورودی، قابل توصیف باشند آنگاه می‌توان گفت که دو هدف مذکور برای مرحله پیش آموزش، محقق گردیده است.

مرحله اول از روش پیشنهادی، برای تغییر یک LSTM تقسیم داده‌ها به اندازه‌هایA دسته واحد است. اندازه دسته A_i وابسته به‌اندازه دیتا بوده و نباید بزرگ‌تر از ۲۰ درصد کل داده‌ها باشد. ما در

این مقاله با تقسیم داده‌ها به دسته‌های مساوی، طول دسته A_i را محاسبه نموده و یک مدل توزیع دوجمله‌ای توصیفی از کل ویژگی‌های متصور برای داده‌های موجود در دسته A_i با کمترین خطای ممکن در پیش آموزش ایجاد خواهیم نمود. مجموع این توزیع‌ها در واقع چندجمله‌ای P متغیره است (P تعداد ویژگی‌های مورد بررسی برای یک داده می‌باشد) که توصیفی از حالت تغییرات داده‌ها بر اساس مدل خود ارائه می‌دهد. در شکل ۲ نمونه‌ای از مقادیر موجود برای داده‌ها را در یک دسته فرضی A_i مشاهده می‌کنید.

	P_1	P_2	P_3	...	P_p
D_1	X_1	--	X_2	...	--
D_2	--	X_3	X_4	...	X_5
...	...	...	...	...	...
D_n	--	X_1	---	...	X_p

شکل ۲: نمایشی از داده‌های موجود در دسته A_i با تاکید بر اینکه این داده‌ها می‌توانند در تعداد ویژگی‌های سازنده یکسان نباشند

سپس یک تابع تغییر حالت با استفاده از رابطه ۱، هر داده موجود در A_i را به صورت زیر رمزگذاری خواهد نمود:

$$y=\sigma (w\tilde{x}+b) \qquad (1)$$

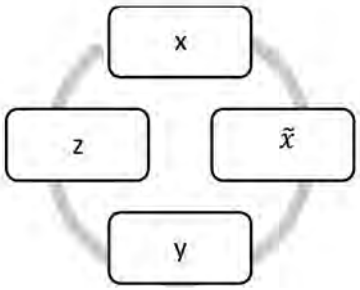
در این رابطه y مقدار تغییر حالت داده شده، w وزن در نظرگرفته شده برای شبکه عصبی بازگشتی در یک سلول (وزن محاسباتی هر bilstm) و b مقدار بایاس آن است. همچنین $\tilde{x}$=|x-Binomial(x)| مقدار حاصل از خطای توزیع دو جمله‌ای است که در مرحله قبل پیش آموزش درLSTM مورد نظر، استفاده شده است.

یک تابع تخمینگر، وظیفه تغییر حالت مقدار y را برای رسیدن به مقدار تخمین مورد نظر در مدل پیشنهادی برای دسته A_i را بر عهده خواهد داشت، این تابع تخمینگر را در رابطه ۲ میتوانید مشاهده نمایید.

$$z=\sigma (\tilde{w}y+\tilde{b}) \qquad (2)$$

در این رابطه مقدار $\tilde{w}$ و $\tilde{b}$ مقادیر وزن و بایاس رابطه تخمینگر خواهند بود. پس از آنکه مقدار z حاصل

شد، این مقدار را با x که داده اصلی است جایگزین نموده و دوباره توزیع دوجمله‌ای ویژگی‌ها را با x جدید خواهیم ساخت. با استفاده از شکل ۳، چرخه جایگزینی این مقادیر را در یک روند به نمایش گذاشته‌ایم.



شکل ۳: نحوه جایگزینی متغیرها در پیش آموزش

این امر موجب تنظیم و به وجود آمدن مدلی در توصیف تغییرات موجود در داده‌ها خواهد شد. برای تنظیم دقیق متغیرها، w را از رابطه ۳ محاسبه خواهیم نمود.

$$w = \frac{-1}{N}\sum_{i=1}^{N}\sum_{j=1}^{P}\log(z_{ij} + (1 - x_{ij}))\log(1 - z_{ij}) \qquad (۳)$$

در واقع، این رابطه، اقتباسی از میزان متوسط تغییر شیب برای دو معادله است. در رابطه ۳، N تعداد داده‌های موردبررسی، میزان zij مقدار داده تخمین زده شده برای داده i در ویژگی‌ام و xij مقدار اصل داده در هر مرحله از انجام آموزش خواهد بود. همچنین مقدار b جدید از رابطه ۴ حاصل می‌گردد:

$$b_{new} = \frac{w_{new}z}{wx+b} \qquad (۴)$$

در واقع w و b دوچند مؤلفه‌ای با تعداد مؤلفه‌های P خواهند بود. این عمل تا زمانی انجام می‌شود که میزان خطای موردانتظار از یک LSTM، از مقدار تعریف شده کمتر باشد.

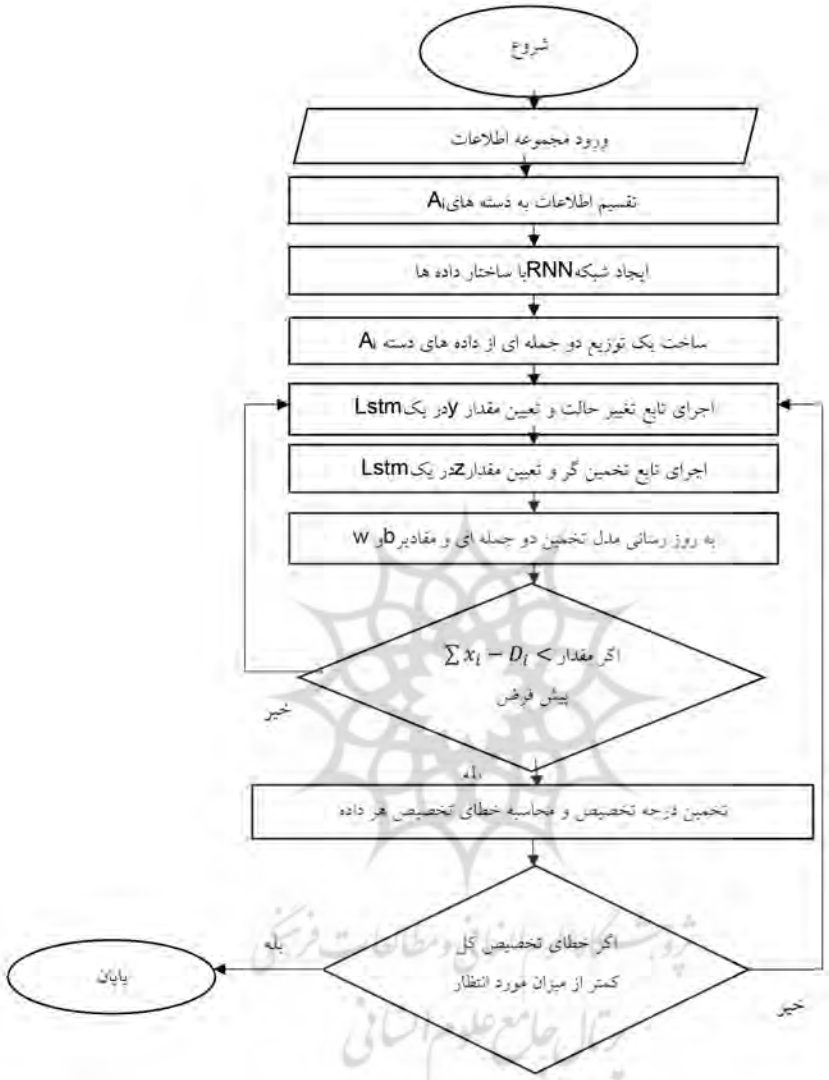
پس از پایان پیش آموزش هر LSTM و برای تخمین میزان نرخ تخصیص داده i ام به دسته j ام میتوان از رابطه ۵ به صورت زیر بهره برد.

$$L_{ij} = \frac{e^{w_i x+b}}{\sum_{j=1}^{N} e^{w_j x+b_j}} \qquad (۵)$$

در این رابطه w_i میزان وزن و b بایاس موجود در شبکه برای داده x در آخرین وضعیت آن، w_j میزان وزن و b_j بایاس موجود برای داده x در هر یک از Lstmهای لایه موردبررسی است. پس از پایان بررسی داده و تخمین نتیجه، میزان E_i خطای تخمین برای هر یک از Lstm ها از طریق رابطه ۶ مجدداً محاسبه می‌گردد.

$$E_i = \sum_{j=1}^{N} \log \frac{L_{jd}}{L_{jr}} \qquad (۶)$$

در این رابطه L_{ir} میزان تخمین برای دستهای که واقعاً داده j عضو آن است و L_{jd} میزان تخمین برای دست‌های است که شبکه تشخیص داده است که داده j ام عضوی از آن دسته است. شبکه تا زمانی به کار خود ادامه خواهد داد که میزان وزن‌های موجود برای هر Lstm بتواند مقدار E_i به بازه متناسب با اهداف کار هدایت نماید. باتوجه‌به مقدار انحراف E_i مقدار w و bهای اولیه مورد استفاده در شبکه تطبیقی موردنظر تنظیم و ویرایش خواهند شد. در واقع اگر بخواهیم روش پیشنهادی را به صورت یک‌روند نمایش دهیم می‌توانیم از شکل ۴ برای این عمل استفاده نمایید.



شکل ٤: شیوه اجرای روش پیشنهادی در بهبود قدرت تفکیک داده‌های نامتوازن

۵- یافته های پژوهش

در این پژوهش برای بررسی مدل پیشنهادی برای پیش بینی و فرضیه‌های خود برای داده‌های نامتوازن، از یک مجموعه داده از سازمان تأمین اجتماعی با موضوع درخواست ازکارافتادگی و تشخیص کمیسیون پزشکی بهره برده‌ایم. تعداد تصادفی ۸۴۶ داده در سال ۱۳۹۹ از معاونت درمان سازمان مربوطه اخذ گردیده است. این داده‌ها شامل داده‌های رشته‌ای و عددی به شرح ذیل می‌باشد:

- داده‌های رشته‌ای: وضعیت ازکارافتادگی، علت مراجعه، جنسیت، نوع بیمه و نتایج کمیسیون‌های قبلی
- داده‌های عددی: شناسه، شناسه بیمه‌شده، کد جنسیت، سن، سابقه بیمه‌پردازی به‌روز، تاریخ شروع طول درمان و تاریخ اتمام طول درمان

داده‌های موجود در مجموعه‌داده پژوهش شامل ۴ ستون از اطلاعات عمومی افراد بیمه‌شده و ۷ ستون از اطلاعات مربوط به سوابق بیمه‌شده، مشخصات بیماری و دیگر داده‌های تخصصی وابسته به سازمان تأمین اجتماعی ایران بوده و نهایتاً یک ستون نتیجه کمیسیون پزشکی و تعیین تکلیف ازکارافتادگی متقاضیان می باشد.

این مجموعه پس از عملیات پیش‌پردازش، حداکثر دارای ۵ ویژگی است، درخواست‌ها در دوکلاس درخواست‌های پذیرفته شده و مردود شده کلاسه‌بندی شده‌اند. با استفاده از این مجموعه داده و بر اساس مطالب ذکر شده، روش پیشنهادی را با مطالعات موجود در منبع (جیانگ و همکاران، ۲۰۱۸) و (یانگ و همکاران، ۲۰۱۸) در ۳ فاکتور دقت تشخیص در پیش‌بینی، زمان مورد نیاز برای پیش‌بینی و میزان مصرف منابع سخت‌افزار مقایسه نموده‌ایم.

دقت تشخیص در این بخش با استفاده از سنجش نرخ تعداد تشخیص صحیح بر تمام داده‌های آزمایش صورت می‌گیرد. لازم به ذکر است که داده‌های آزمایش از میزان ۲۰ درصد کل داده‌های پروژه به صورت تصادفی و با استراتژی k-fold انتخاب میگردند. برای از بین بردن اثر داده‌ها ۱۰ بار روش‌های مورد مطالعه اجرا شده و میانگین نتایج کل اجراها برای نتایج نهایی انتخاب شدهاند. واحد سنجش زمان مورد نیاز برای پردازش ثانیه و میزان مصرف منابع پردازشی برحسب میزان درگیر بودن CPU در فرآیند پردازش برحسب واحد زمان/ واحد پردازش و میزان منابع حافظهای مصرفی از حافظه RAM برحسب KB محاسبه خواهد شد.

داده‌های مورد آزمایش در یک CPU با ۴ واحد پردازشی و با حداکثر دامنه پردازشی ۲,۲ مگاهرتز پردازش شده‌اند. همچنین محیط شبیه‌سازی در یک نرم‌افزار شبیه‌سازی MATLAB ۲۰۱۸ ایجاد شده است.

تعداد کل LSTMهای ایجاد شده بیش از ۷۱۵ هزار LSTM است که برابر با کل واحدهای عددی

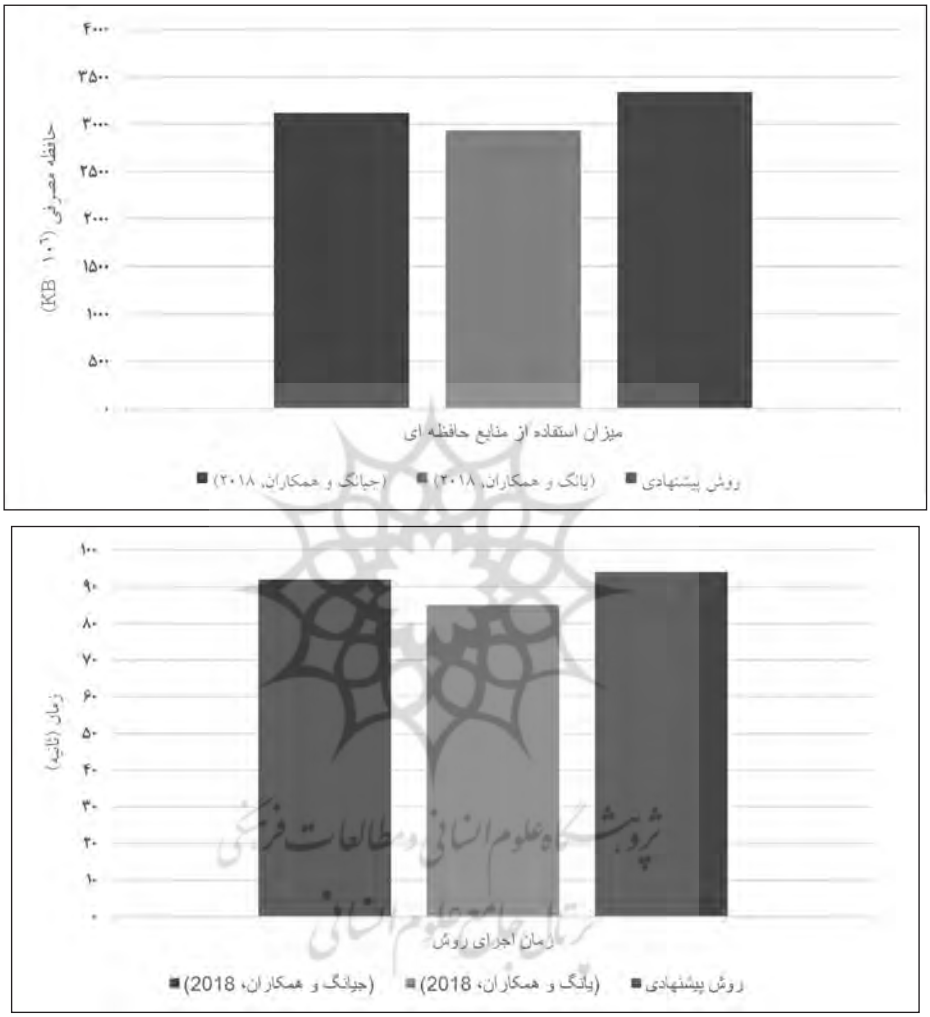
ورودی میباشد. هر LSTM دارای ۲ خروجی و ۳ ورودی میباشد.

هر آزمایش برای سنجش هر نقطه از محور هریک از نمودارهای مندرج در این بخش به تعداد ۱۰ مرتبه انجام شده و پس از حذف بهترین و بدترین نتیجه، میزان متوسط نتایج گزارش شده در جداول ثبت گردیده است. همچنین علاوه بر موارد ذکر شده، میزان پیش فرض برخی از متغیرهای تأثیرگذار در اجرای شبیه‌سازی، در جدول ۱ آمده است.

جدول ۱. مقادیر متغیرهای مورد استفاده با مقدار دهی اولیه

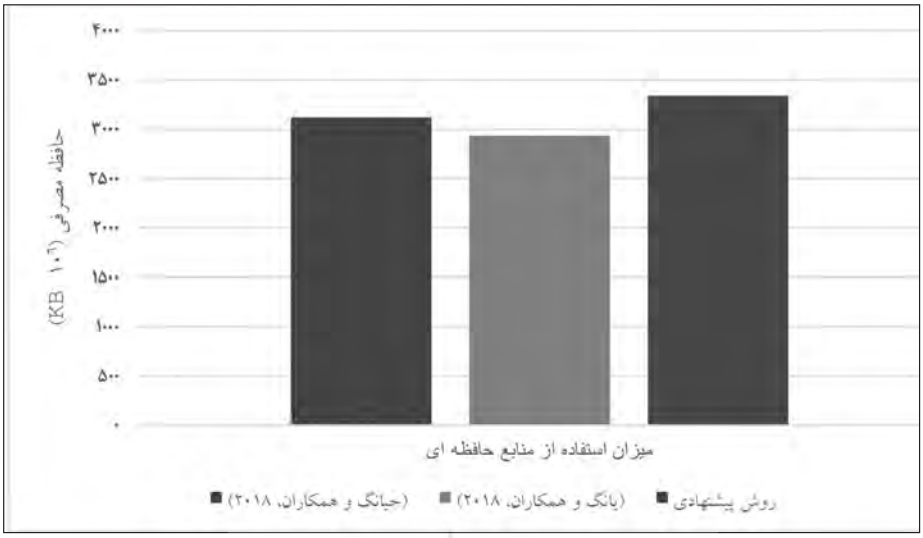
ردیف	عنوان متغیر	شرح	میزان پیش‌فرض
۱	μ	تنظیم‌کننده دوره افزایش یا کاهش	۰,۰۱
۲	w	وزن محاسباتی هر biLSTM	مقدار اولیه ۱
۳	b	مقدار بایاس هر biLSTM	مقدار اولیه تصادفی
۴	C_t	وضعیت خروجی/ورودی	مقدار اولیه ۱
۵	h_t	خروجی/ورودی	مقدار اولیه ۱
۶	R_t	حداکثر تعداد مقایسه در یک مرحله برای به دست آوردن بهترین C_t	۱۰۰ مرتبه
۷	W	میزان وزن هرکدام از فاکتورهای هزینه، منابع و دقت	۰,۳۳
۸	حذف واحد	حداکثر حذف واحدها از لایه در شروع پردازش	۱۰
۹	واحد لایه‌های پنهان	حداکثر واحد موجود در لایههای پنهان	۱۰۰
۱۰	تعداد تکرار	حداکثر تعداد تکرار بدون تغییر در biLstm	۱۰
۱۱	Epouch	بیشترین Epouch	۱۰۰
۱۲	batch size	حداقل	۴

پس از اجرای شبیه سازی با شرایط ذکر شده میزان زمان مصرفی هر یک از روش‌های مورد مطالعه به صورت زیر محاسبه شده است.



شکل ۵: میزان زمان مصرفی در ۳ روش مورد مقایسه

همچنین میزان منابع مصرفی مورد استفاده در روش پیشنهادی و ۲ روش مورد مقایسه به ترتیب از نظر میزان حافظه و میزان استفاده از واحد پردازشگر مرکزی به ترتیب در شکل‌های ۶ و ۷ برای مقایسه به نمایش گذاشته شده‌اند.

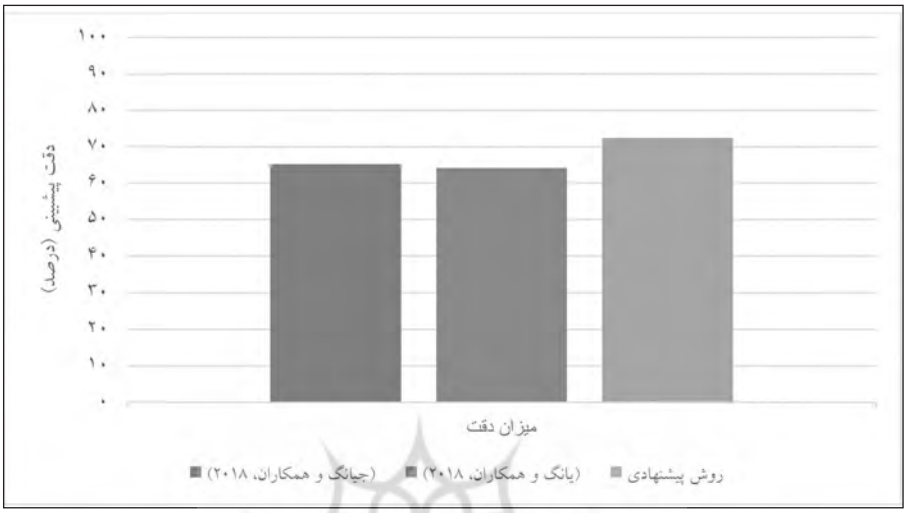


شکل ۶: میزان استفاده از منابع حافظه‌ای (RAM)

شکل ۷: زمان استفاده از CPU

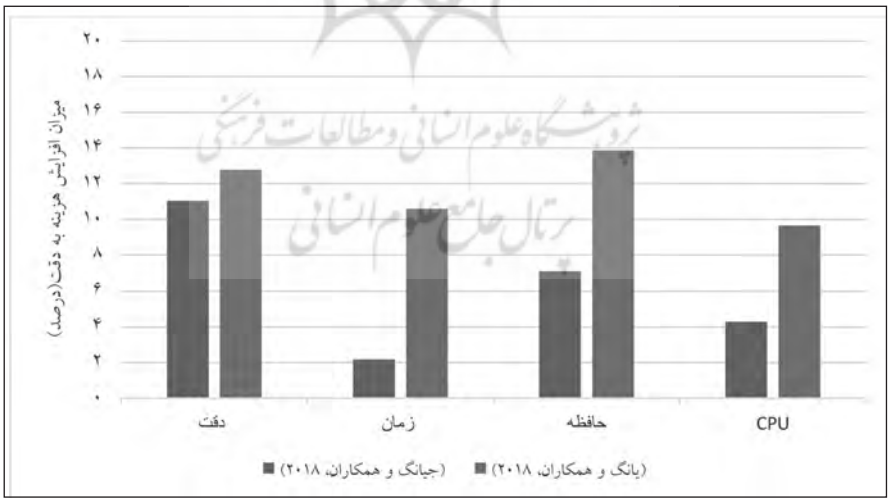
اما می‌خواهیم نشان دهیم تغییرات مثبت روند پیشنهادی در میزان دقت نتیجه‌گیری به چه اندازه است. میزان متغیر دقت در پیش‌بینی کلاس‌های هر نمونه طبق سناریوی پیشنهادی در ۳ روش مورد

مطالعه نیز در نمودار شکل ۸ قابل مشاهده است.



شکل ۸: میزان دقت در روش‌های مورد مطالعه

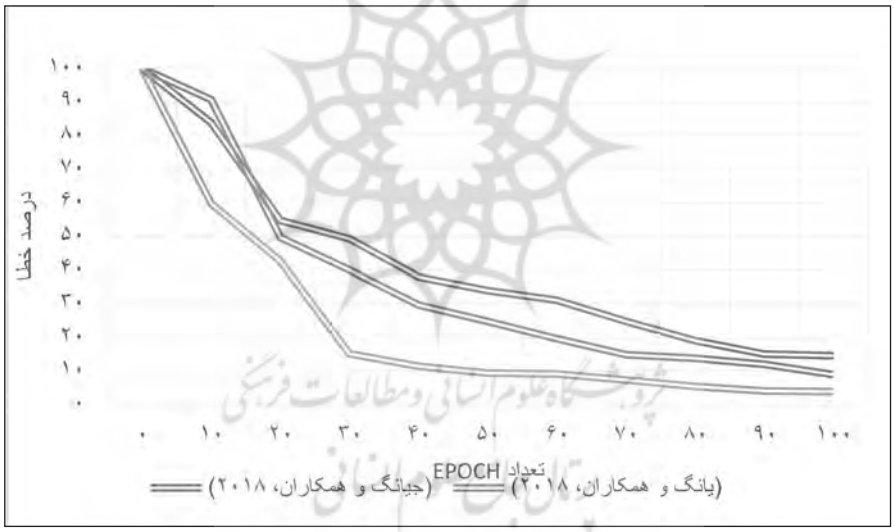
مقایسه کلی از میزان افزایش هزینه نسبت به افزایش کارایی روش پیشنهادی و مدل‌های مقایسه‌ای به نسبت درصدی در شکل ۹ نشان داده شده است.



شکل ۹. نسبت افزایش دقت در مقابل افزایش منابع در روش پیشنهادی و در مقایسه با روش‌های (جیانگ و همکاران، ۲۰۱۸) و (یانگ و همکاران، ۲۰۱۸)

در واقع به زبان ساده‌تر به طور متوسط با افزایش ۴,۵ درصدی هزینه‌ها (CPU، حافظه و زمان) نسبت به مقاله (جیانگ و همکاران، ۲۰۱۸) میزان ۱۱,۰۴ درصد، دقت پیش‌بینی کلاس‌ها در مدل پیشنهادی افزایش پیدا خواهد کرد. این رقم در مقایسه با مقاله (یانگ و همکاران، ۲۰۱۸) در مقابل افزایش ۱۱,۳ درصد افزایش منابع، دقت را به اندازه ۱۲,۷۷ درصد افزایش داده است.

میزان خطای آموزش در طول زمان برای ۱۰۰ epoch را می‌توان در نمودار شکل ۱۰ به نمایش گذاشت. این نمودار نشان می‌دهد نسبت به تعداد بسته‌های داده شرکت‌کننده در آموزش روند کاهش خطا در روش پیشنهادی در تعداد داده‌های کمتری به وقوع می‌پیوندد. این روند حرکت روش پیشنهادی به سمت کاهش میزان خطا با وجود داده‌های نامتوازن نشان می‌دهد روش پیشنهادی نیازمند داده‌های کمتری نسبت به دو روش ذکر شده است، به عبارتی عملکرد و انعطاف بهتری نسبت به دو روش (جیانگ و همکاران، ۲۰۱۸) و (یانگ و همکاران، ۲۰۱۸) در اندازه داده‌های مختلف در مرحله آموزش دارد.



شکل ۱۰: روند تغییرات میزان خطای آموزش نسبت به تعداد داده‌ها

۶- بحث و نتیجه گیری

در این مطالعه، یک چهارچوب نوآورانه مبتنی بر یادگیری عمیق بازگشتی با استفاده از شبکه‌های حافظه بلند- کوتاه مدت (LSTM) برای پیش‌بینی نتایج فرآیندهای بیمه‌های اجتماعی ارائه شد.

در این نوع از شبکه چند لایه، هر سطح به صورت برخط، هم با گرههای لایه L و هم با گره‌های لایه L+۱ در ارتباط بوده و هم‌زمان با دریافت هر ورودی، پردازش می‌کند و خروجی جدید تحویل خواهد داد و همچون تمامی شبکه‌های عصبی بازگشتی دارای یک وضعیت داخلی است و در هر مرحله برای تصمیم‌گیری، علاوه بر ورودی جدید به آن وضعیت هم رجوع می‌کند که با هر ورودی جدید به‌روزرسانی خواهد شد.

از آنجایی‌که داده‌های ورودی از بیمه‌های اجتماعی دارای عدم توازن در ساختار خود بوده، قبل از آنکه داده‌ها وارد روند آموزش گردند، یک مرحله پیش آموزش روی داده‌ها انجام گرفته و این رخداد (پیش آموزش) در هر LSTM، روی داده‌ها صورت گرفته است.

برای نشان دادن عملکرد هر روش، مقایسه آن با روش‌های قبلی، امری متداول بوده؛ بنابراین نتایج حاصل از این تحقیق، با دو روش معتبر مقایسه گردیده است. استفاده از روش ارائه شده، میزان استفاده از منابع مصرفی را به طور نسبی افزایش داده؛ ولی میزان خطا با کاهش چشمگیری مواجه شده است. همچنین میزان خطای آموزش در یادگیری عمیق بازگشتی نسبت به تعداد داده‌های یکسان، کاهش یافته و نسبت به روش‌های مشابه دیگر، در اجرا، عملکرد و اخذ نتایج دارای شرایط مطلوب‌تری می‌باشد.

این مدل توانسته است به طور موفقیت‌آمیزی به مهم‌ترین چالش‌های داده‌های بیمه‌های اجتماعی شامل حجم بالای ویژگی‌ها، ساختار ناهمگن داده‌ها، توزیع نامتعادل کلاس‌ها و کمبود نمونه‌های داده‌ای پاسخ دهد.

نتایج حاصل از آزمایش‌های گسترده روی داده‌های واقعی سازمان تأمین اجتماعی ایران، اثربخشی قابل توجه مدل پیشنهادی را در افزایش دقت پیش‌بینی در کنار مصرف منطقی منابع محاسباتی نشان می‌دهد. این دستاورد، نشان‌دهنده برتری عملکرد مدل نسبت به روش‌های مرجع پیشین است و قابلیت تطبیق و یادگیری بهتر از داده‌های نامتوازن را به اثبات رسانده است. به ویژه، سرعت همگرایی مدل در طول فرایند آموزش و کاهش خطای آن با تعداد نمونه‌های کمتر، بیانگر انعطاف‌پذیری بالا و کاربردی بودن این رویکرد در محیط‌های عملیاتی واقعی است.

مدل پیشنهادی در این تحقیق با استفاده از شبکه‌های LSTM چندلایه و تکنیک‌های پیش‌پردازش ویژه، دقت پیش‌بینی را به میزان ۱۱ تا ۱۲ درصد نسبت به روش‌های مطرح در (جیانگ و همکاران، ۲۰۱۸) و (یانگ و همکاران، ۲۰۱۸) بهبود بخشیده است. این بهبود قابل توجه، نمایانگر قدرت

یادگیری عمیق بازگشتی در استخراج الگوهای پنهان از داده‌های نامتوازن و ناهمگن است که در تحقیقات پیشین کمتر به این میزان دست یافته‌اند.

برخلاف روش‌های سنتی که برای رسیدن به دقت مطلوب نیازمند حجم بالایی از داده‌های آموزش هستند، مدل حاضر توانسته است با داده‌های کمتر و با کارایی بالاتر، نرخ خطای پایین‌تری داشته باشد. این موضوع نشان می‌دهد که روش پیشنهادی از انعطاف‌پذیری و قابلیت تعمیم‌پذیری بیشتری برخوردار است.

بسیاری از مطالعات پیشین به چالش‌های داده‌های ناهمگن و توزیع نامتوازن کمتر پرداخته‌اند یا از روش‌های ساده‌تری استفاده کرده‌اند که در شرایط واقعی کاربرد محدودی دارند. درحالی‌که مدل پیشنهادی با استفاده از مرحله پیش آموزش و تقسیم‌بندی داده‌ها به دسته‌های مشخص، این چالش‌ها را به طور مؤثرتری مدیریت کرده است.

از سوی دیگر، طراحی ساختار چندلایه و پردازش برخط مدل، امکان به‌روزرسانی مداوم و تصمیم‌گیری پویا را فراهم می‌آورد که در کاربردهای زمان واقعی مانند پیش‌بینی وضعیت درخواست‌های بیمه‌ای، بسیار حیاتی است. مرحله پیش‌پردازش و پیش آموزش اختصاصی هر واحد LSTM نیز به بهبود توانایی مدل در مدیریت ناهمگنی و نابرابری داده‌ها کمک کرده است، موضوعی که در بسیاری از سیستم‌های سنتی یادگیری ماشینی کمتر به آن توجه شده است.

باتوجه‌به چشم‌انداز توسعه سامانه‌های هوشمند در سازمان‌های بیمه‌ای، مدل پیشنهادی قابلیت گسترش و پیاده‌سازی در مقیاس وسیع‌تر را دارد و می‌تواند به عنوان ابزاری کلیدی در بهبود تصمیم‌گیری‌های استراتژیک، افزایش بهره‌وری، کاهش هزینه‌ها و ارتقای کیفیت خدمات بیمه‌ای مورد استفاده قرار گیرد. همچنین، این روش می‌تواند بستری برای پژوهش‌های آتی در حوزه پیش‌بینی داده‌های نامتوازن و ناهمگن فراهم آورد و به توسعه راهکارهای هوشمند در سایر بخش‌های مرتبط با داده‌های پیچیده و بزرگ کمک کند.

باتوجه‌به مطالب و نتایج گفته شده در این تحقیق، پیشنهاد برای کارهای آتی به شرح ذیل ارائه می‌گردد

- در خصوص ایجاد مدلی که بتواند بابت داده‌های با نتایجی غیر از صفر و یک نیز پیش بینی داشته باشد و برای فرایندهای مختلف در زمینه‌های گوناگون سازمان تأمین اجتماعی، نتایج قابل‌قبولی ارائه دهد، می‌توان تحقیق کرد.
- مدلی که صرفاً با کسب داده‌های مختلف کسب‌وکار و فارغ از برچسب‌هایی که خواهند داشت، بتواند الگوی مناسب را ساخته و پیش‌بینی‌های مربوطه را انجام دهد نیز از اهداف بعدی می‌باشد.

- باتوجه‌به چشم‌انداز سازمان تأمین اجتماعی و روند هوشمندسازی فرایندهای سازمان مذکور، دقت پیش‌بینی در فرایندهای کسب‌وکاری دارای اهمیت فراوانی بوده؛ لذا استفاده از سایر روش‌های هوش مصنوعی جهت افزایش و بهبود دقت، بسیار کارآمد خواهد بود.

منابع

1. Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends® in Machine Learning, 2*(1), 1–127.
2. Bevacqua, A., Carnuccio, M., Folino, F., Guarascio, M., & Pontieri, L. (2014). A data-driven prediction framework for analyzing and monitoring business process performances. In S. Hammoudi, J. Cordeiro, L. A. Maciaszek, & J. Filipe (Eds.) , *ICEIS 2013. LNBIP* (Vol. 190, pp. 100–117). Springer, Cham. https://doi.org/10.1007/978-3-319-09492-2_7
3. Bolt, A., & Sepúlveda, M. (2014). Process remaining time prediction using query catalogs. In N. Lohmann, M. Song, & P. Wohed (Eds.) , *BPM 2013. LNBIP* (Vol. 171, pp. 54–65). Springer, Cham. https://doi.org/10.1007/978-3-319-06257-0_5
4. Briot, J.-P., Hadjeres, G., & Pachet, F. (2017). Deep learning techniques for music generation—a survey. *arXiv preprint* arXiv:1709.01620.
5. Calheiros, R. N., Masoumi, E., Ranjan, R., & Buyya, R. (2015). Workload prediction using ARIMA model and its impact on cloud applications' QoS. *IEEE Transactions on Cloud Computing, 3*, 449–458.
6. Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th International Conference on Machine Learning* (pp. 160–167). ACM.
7. Emmert-Streib, F., de Matos Simoes, R., Glazko, G., McDade, S., Haibe-Kains, B., Holzinger, A., Dehmer, M., & Campbell, F. C. (2014). Functional and genetic analysis of the colon cancer network. *BMC Bioinformatics, 15*(6), S6.
8. Evermann, J., Rehse, J.-R., & Fettke, P. (2016). A deep learning approach for predicting process behavior at runtime. In *International Conference on Business Process Management* (pp. 327–338). Springer.
9. Farasat, A., Nikolaev, A., Srihari, S. N., & Blair, R. H. (2015). Probabilistic graphical models in modern social network analysis. *Social Network Analysis and Mining, 5*, 62.
10. Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research, 270*(2), 654–669.
11. Folino, F., Guarascio, M., & Pontieri, L. (2012). Discovering context-aware models for predicting business process performances. In R. Meersman et al. (Eds.) , *OTM 2012. LNCS* (Vol. 7565, pp. 287–304). Springer. https://doi.org/10.1007/978-3-642-33606-5_18
12. Folino, F., Guarascio, M., & Pontieri, L. (2013). Context-aware predictions on business processes: An ensemble-based solution. In A. Appice, M. Ceci, C. Loglisci, G. Manco, E. Masciari, & Z. W. Ras (Eds.) , *NFMCP 2012. LNCS (LNAI)* (Vol. 7765, pp. 215–229). Springer. https://doi.org/10.1007/978-3-642-37382-4_15

13. Geng, R., Bose, I., & Chen, X. (2015). Prediction of financial distress: An empirical study of listed Chinese companies using data mining. *European Journal of Operational Research, 241*(1), 236–247.

14. Gibson, T. A., Henderson, J. A., & Wiles, J. (2014). Predicting temporal sequences using an event-based spiking neural network incorporating learnable delays. (pp. 3213–3220).

15. Globerson, A., Chechik, G., Pereira, F., & Tishby, N. (2006). Embedding heterogeneous data using statistical models. In *Proceedings of The National Conference on Artificial Intelligence* (Vol. 21, No. 2, p. 1605).

16. Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*.

17. He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284.

18. Heaton, J. B., Polson, N. G., & Witte, J. H. (2017). Deep learning for finance: Deep portfolios. *Applied Stochastic Models in Business and Industry, 33*(1), 3–12.

19. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*, 1735–1780.

20. Hu, R., Yu, C. P., Fung, S.-F., Pan, S., Wang, H., & Long, G. (2017). Universal network representation for heterogeneous information networks. In *2017 International Joint Conference on Neural Networks (IJCNN)* (pp. 388–395). IEEE.

21. Huang, X., & An, A. (2002). Discovery of interesting association rules from livelink web log data. In *Data Mining. ICDM 2003. Proceedings. 2002 IEEE International Conference on* (pp. 763–766). IEEE.

22. Huck, N. (2009). Pairs selection and outranking: An application to the S&P 100 index. *European Journal of Operational Research, 196*(2), 819–825.

23. Jiang, X. R., & Le, G. (2005). Microarray gene expression data association rules mining based on BSC-tree and S-tree. *Data & Knowledge Engineering, 53*, 3–29.

24. Jiang, X., Pan, S., Long, G., Xiong, F., Jiang, J., & Zhang, C. (2018). Cost-sensitive parallel learning framework for insurance intelligence operation. *Transactions on Industrial Electronics*, 1–11.

25. Jozefowicz, R., Zaremba, W., & Sutskever, I. (2015). An empirical exploration of recurrent network architectures. In *International Conference on Machine Learning* (pp. 2342–2350).

26. Khan, A., Le, H., Do, K., Tran, T., Ghose, A., Dam, H., & Sindhgatta, R. (2018). Memory-augmented neural networks for predictive process analytics. *arXiv preprint* arXiv:1802.00938.

27. Kraus, M., & Feuerriegel, S. (2017). Decision support from financial disclosures with deep neural networks and transfer learning. *Decision Support Systems, 104*, 38–48.

28. Kumar, J., & Singh, A. K. (2018). Workload prediction in cloud using artificial neural network

and adaptive differential evolution. *Future Generation Computer Systems, 81,* 41–52.

29. Lakshmanan, G. T., Shamsi, D., Doganata, Y. N., Unuvar, M., & Khalaf, R. (2015). A Markov prediction model for data-driven semi-structured business processes. *Knowledge and Information Systems, 42,* 97–126.

30. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

31. Lessmann, S., Baesens, B., Seow, H.-V., & Thomas, L. C. (2015). Benchmarking state-of-the-art classification algorithms for credit scoring: An update of research. *European Journal of Operational Research, 247*(1), 124–136.

32. Letham, B., Rudin, C., & Madigan, D. (2013). Sequential event prediction. *Machine Learning, 93,* 357–380.

33. Lin, L., Wen, L., & Wang, J. (2019). MM-Pred: A deep predictive model for multi-attribute event sequence. In *Proceedings of the 2019 SIAM International Conference on Data Mining* (pp. 118–126). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611975673

34. Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing, 234,* 11–26.

35. Márquez-Chamorro, A. E., Resinas, M., & Ruiz-Corts, A. (2017). Predictive monitoring of business processes: A survey. *IEEE Transactions on Services Computing, PP*(99), 1–1. https://doi.org/10.1109/TSC.2017.2772256

36. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).

37. Montufar, G. F., Pascanu, R., Cho, K., & Bengio, Y. (2014). On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems* (pp. 2924–2932).

38. Niimi, J., & Hoshino, T. (2017). Predicting purchases using the variety of customer behaviors: Analysis of the purchase history and the browsing history by deep learning. *Transactions of the Japanese Society for Artificial Intelligence, 32.*

39. Oztekin, A., Kizilaslan, R., Freund, S., & Iseri, A. (2016). A data analytic approach to forecasting daily stock returns in an emerging market. *European Journal of Operational Research, 253*(3), 697–710.

40. Pan, S., Hu, R., Long, G., Jiang, J., Yao, L., & Zhang, C. (2018). Adversarially regularized graph autoencoder. *arXiv preprint* arXiv:1802.04407.

41. Pan, S., Wu, J., Zhu, X., Zhang, C., & Wang, Y. (2016). Tri-party deep network representation. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (pp. 1895–1901).

42. Pandey, S., Nepal, S., & Chen, S. (2011). A test-bed for the evaluation of business process pre-

diction techniques. In *7th International Conference on Collaborative Computing: Networking, Applications and Worksharing, CollaborateCom 2011* (pp. 382–391). Orlando, FL, USA.

43. Pavlidis, P., Weston, J., Cai, J., & Grundy, W. N. (2001). Gene functional classification from heterogeneous data. In *Proceedings of the Fifth Annual International Conference on Computational Biology* (pp. 249–255). ACM.

44. Pfeffer, A. (2005). Functional specification of probabilistic process models. In *AAAI* (pp. 663–669).

45. Polato, M., Sperduti, A., Burattin, A., & de Leoni, M. (2014). Data-aware remaining time prediction of business process instances. In *2014 International Joint Conference on Neural Networks, IJCNN 2014*, Beijing, China, July 6–11, 2014 (pp. 816–823).

46. Polato, M., Sperduti, A., Burattin, A., & de Leoni, M. (2016). Time and activity sequence prediction of business process instances. *CoRR*, abs/1602.07566.

47. Provost, F. (2000). Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI 2000 Workshop on Imbalanced Data Sets* (pp. 1–3).

48. Rogge-Solti, A., & Weske, M. (2013). Prediction of remaining service execution time using stochastic Petri nets with arbitrary firing delays. In S. Basu, C. Pautasso, L. Zhang, & X. Fu (Eds.) , *ICSOC 2013. LNCS* (Vol. 8274, pp. 389–403). Springer. https://doi.org/10.1007/978-3-642-45005-1_27

49. Rogge-Solti, A., & Weske, M. (2015). Prediction of business process durations using non-Markovian stochastic Petri nets. *Information Systems, 54,* 1–14.

50. Rudolph, M., Ruiz, F., Mandt, S., & Blei, D. (2016). Exponential family embeddings. In *Advances in Neural Information Processing Systems* (pp. 478–486).

51. Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks, 61*, 85–117.

52. Schwegmann, B., Matzner, M., & Janiesch, C. (2013). preCEP: Facilitating predictive event-driven process analytics. In J. Brocke, R. Hekkala, S. Ram, & M. Rossi (Eds.) , *DESRIST 2013. LNCS* (Vol. 7939, pp. 448–455). Springer. https://doi.org/10.1007/978-3-642-38827-9_36

53. Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., & Poggio, T. (2007). A quantitative theory of immediate visual recognition. *Progress in Brain Research, 165*, 33–56.

54. Sirignano, J. (n.d.). Deep learning for limit order books. *CoRR*. abs/1601. URL https://arxiv.org/abs/1601

55. Sirignano, J. A., Sadhwani, A., & Giesecke, K. (2016). Deep learning for mortgage risk. URL https://people.stanford.edu/giesecke/

56. Tax, N., Verenich, I., La Rosa, M., & Dumas, M. (2017). Predictive business process monitoring with LSTM neural networks. In *International Conference on Advanced Information Systems*

Engineering (pp. 477–492). Springer.

57. Van der Aalst, W. M. P., Schonenberg, M. H., & Song, M. (2011). Time prediction based on process mining. *Information Systems, 36*(2), 450–475.

58. Van Dongen, B. F., Crooy, R. A., & Aalst, W. M. P. (2008). Cycle time prediction: When will this case finally be finished? In R. Meersman & Z. Tari (Eds.) , *OTM 2008. LNCS* (Vol. 5331, pp. 319–336). Springer. https://doi.org/10.1007/978-3-540-88871-0_22

59. Wang, C., Pan, S., Long, G., Zhu, X., & Jiang, J. (2017). MGAE: Marginalized graph autoencoder for graph clustering. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management* (pp. 889–898). ACM.

60. Wang, J., Yang, Y., Mao, J., Huang, Z., Huang, C., & Xu, W. (2016). CNN-RNN: A unified framework for multi-label image classification. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2285–2294). IEEE.

61. Wang, X., & Wang, Y. (2014). Improving content-based and hybrid music recommendation using deep learning. In *Proceedings of the 22nd ACM International Conference on Multimedia* (pp. 627–636). ACM.

62. Yang, Y., Kolesnikova, A., Lessmann, S., Ma, T., Sung, M.-C., & Johnson, J. E. V. (2018). Can deep learning predict risky retail investors? A case study in financial risk behavior forecasting. *arXiv*. https://arxiv.org/abs/1812.06175

63. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016). Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1480–1489).