پیش‌بینی بقای بیماران مبتلا به سرطان پستان با استفاده از بهینه‌سازی مدل DeepHit

سهیلا رضایی، حسین قیومی‌زاده*، محمدحسین قلی‌زاده و علی فیاضی
گروه مهندسی برق، دانشکده فنی و مهندسی، دانشگاه ولی‌عصر (عج) رفسنجان، رفسنجان، ایران.

چکیده

با توجه به اهمیت و شیوع سرطان پستان به‌عنوان دومین علت مرگ در بین بیماری‌های سرطانی در جهان، دسترسی به مدل‌هایی را که با دقت بالا بتوانند بقای این بیماران در افراد مبتلا پیش‌بینی کنند، موردتوجه است. هدف از این مطالعه، استفاده از شبکه عصبی عمیقِ بهینه‌سازی‌شده برای پیش‌بینی بقای بیماران مبتلا به سرطان پستان است. مطالعهٔ حاضر یک مطالعه تحلیلی است. داده‌های مورداستفاده از بانک داده‌ای METABRIC مربوط به طبقه‌بندی مولکولی از بیماران مبتلا به سرطان سینه مجمع بین‌المللی است. تعداد کل بیماران موردبررسی ۱۹۸۱نفر است؛ از این تعداد ۸۸۸ نفر از بیماران تا لحظه مرگ تحت مراقبت و بقیه در حین مطالعه از ادامه مطالعه صرف‌نظر کرده‌اند. در این مجموعه‌دادگان به ۲۱ ویژگی بالینی بیماران توجه شده است که در کل شامل شش ویژگی کمی و پانزده ویژگی کیفی است. جهت پیش‌بینی بقا از مدل شبکه عصبی عمیق DeepHit بهینه‌سازی‌شده استفاده می‌شود. مدل بهینه‌سازی شده توانسته‌است معیار c_index=۰/۷۳ را، که معیاری برای سنجش قابلیت مدل‌های آنالیز بقا است کسب کند. مقایسه با مدل‌های قبلی بر اساس مجموعه‌داده‌های واقعی و مصنوعی نشان می‌دهد که DeepHit بهینه‌سازی‌شده به پیشرفت‌های عملکردی بزرگ و آماری قابل‌توجهی نسبت به روش‌های سطح بالا دست یافته‌است.

واژگان کلیدی: استخراج ویژگی، یادگیری عمیق، تجزیه و تحلیل بقا، سرطان پستان.

Predicting the Survival of Breast Cancer Patients using DeepHit model optimization

Soheila Rezaei, Hossein Ghayoumi Zadeh*, Mohammad Hossein Gholizadeh and Ali Fayazi
Dept. of Electrical Engineering, Faculty of Engineering, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran.

Abstract

Predicting and estimating the time it takes for an event of interest to occur base on available information is special assistance in how to deal with the event and handle it or provide solution to prevent the occurrence of the event. In medicine, valuable information about evaluating the types of treatments and prognosis and providing solution to handle event can be gained by predicting the time that an event occurrence according to information recorded from patients. Many statistical solutions have been proposed for predicting the time that an event occurrence and the most professional method is Survival Analysis. The purpose of Survival Analysis is to predict the time that an event occurrence a model effective parameters in estimating the time, which can be control or eliminating problematic factors. Due to the importance and prevalence of breast cancer as the second leading cause of death among cancer patients in the world, access to models that can accurately predict the survival of breast cancer patients is very important. The present study is an analytical study. The data used in this study are

* نویسندهٔ عهده‌دار مکاتبات
* Corresponding author

• تاریخ ارسال مقاله: ۱۴۰۰/۱۲/۲۶ • تاریخ پذیرش: ۱۴۰۲/۳/۱۲ • تاریخ انتشار: ۱۴۰۲/۱۰/۲۴ • نوع مطالعه: پژوهشی
سال ۱۴۰۲ شمارهٔ ۳ پیاپی ۵۷

[DOI: 10.61186/jsdp.20.3.3]

taken from The Molecular Taxonomy Data of the International Federation of Breast Cancer (METABRIC) database, which is related to which is related to the molecular classification of breast cancer patients. The total number of patients studied was 1981. Of these, 888 patients were in care until the time of death and the rest did not continue the study during the study. In this database, 21 clinical features of patients have been considered, which includes a total of 6 quantitative features and 15 qualitative features. To predict survival, a deep neural network model called the optimized DeepHit is used. The optimized model has achieved the criterion of c_index = 0.73, which is a criterion for measuring the capability of survival analysis models. Comparisons with previous models based on real and synthetic datasets show that the optimized DeepHit has achieved great performance and statistically significant improvements over previous advanced methods.

Keywords: Feature extraction, Deep learning, Survival Analysis, Breast cancer

۱- مقدمه

سرطان شرایطی است که در آن سلول‌ها به‌طور نامنظم رشد می‌کنند و بر دیگر بخش‌های بدن تأثیر می‌گذارند [۱]. سرطان سینه با داشتن یک‌میلیون مورد جدید در جهان در سال، شایع‌ترین بدخیمی در زنان است و هیجده درصد از کل سرطان‌های زنان را شامل می‌شود. در انگلستان، که در آن شیوع استاندارد سنی و مرگ‌ومیر در جهان بالاترین است، شیوع در میان زنان پنجاه ساله نزدیک به دو در هزار زن در سال است، و این بیماری شایع‌ترین علت مرگ‌ومیر در میان زنان چهل تا پنجاه ساله است، که حدود یک‌پنجم کل مرگ‌ومیر در این گروه سنی را شامل می‌شود [۳,۲].

ازآنجایی که روش‌های مناسب‌تر غربال‌گری و پیش‌گیری از وقوع این بیماری و همچنین روش‌های تشخیصی و درمانی جدید و دقیق‌تری برای سرطان پستان به‌وجود آمده است، به نظر می‌رسد که در آینده نزدیک و در سال‌های آینده میزان مرگ‌ومیر ناشی از این بیماری به‌خصوص در کشورهای توسعه‌یافته کاهش چشم‌گیری داشته و میزان بقا در این بیماری رو به افزایش باشد [۴]. اگرچه سرطان پستان دومین علت عمده مرگ ناشی از سرطان در زنان است، اما میزان بقای آن زیاد است و با تشخیص زودرس و ارائه راهکارهای ویژه می‌توان منجر به کاهش مراجعه دیرهنگام، ارائه درمان مؤثر به‌منظور افزایش بقاء، کاهش مرگ و ارتقای کیفیت زندگی بیماران شد. به‌این‌ترتیب ۹۷ درصد از زنان حداقل ۵ سال زنده می‌مانند [۵].

هیچ علت اصلی و اختصاصی برای سرطان پستان وجود ندارد و در ایجاد آن ترکیبی از عوامل مختلفی وجود دارد. از عواملی که با کاهش بقا در این بیماری در ارتباط هستند می‌توان به مراحل بالاتر بیماری، سن بالا، افزایش تعداد غدد لنفاوی درگیر، افزایش شدت بالاتر تومور، بیان گیرنده‌های منفی استروژنی و پروژسترونی، بیان بالای انکوژنهایی مانند فاکتور رشد اپیدرمی انسانی Her2neu، دریافت انواع درمان‌ها مانند جراحی، رادیوتراپی، شیمی‌درمانی، وضعیت اقتصادی- اجتماعی پایین مانند تحصیلات پایین اشاره کرد [۶]. با توجه به گسترش روزافزون سرطان، تحقیقات گسترده‌ای برای کشف راه‌های پیشگیری و درمان آن و شناسایی داروهای مؤثر در این زمینه انجام می‌شود.

در سال‌های اخیر استفاده از نشانگرهای زیستی و مولکولی به دلیل حساسیت و اختصاصیت بالا اهمیت زیادی پیداکرده است [۷]. نشانگرهای تومور از مهم‌ترین نشانگرهای مولکولی هستند که به‌نوعی در بروز و پیشروی سرطان دخیل بوده و به‌طور مستقیم توسط تومور یا توسط سلول‌های طبیعی در اثر پاسخ به حضور تومور تولید می‌شوند. پیشرفت در تکنیک‌های مولکولی با توان بالا و بیوانفورماتیک به درک بهتر زیست‌شناسی سرطان و توسعه روش‌های مولکولی پیش‌آگهی دهنده و پیش‌بینی کننده جدید، کمک کرده است [۷]. پارامترهای تشخیص مورفولوژیکی سرطان پستان شامل اندازه و درجه تومور و وضعیت مثبت یا منفی بودن نشانگرهای ایمونوهیستوشیمی نظیر گیرنده‌های استروژن، گیرنده‌های پروژسترون، HER2 و نشانگر Ki68 است. از این دیدگاه، انواع تومورهای پستان به پنج دسته اصلی با علائم بالینی متمایز تقسیم می‌شوند که عبارت‌اند از Luminal A ، Luminal B، HER2 Overexpression، Normal-Like ، Basal-Like گروه‌های Luminal نشانگرهای متمایز ER و PR، HER2 را نشان می‌دهند، ولی در گروه Like-Basal هر سه منفی هستند [۸, ۹].

تحلیل داده، فرآیند ارزیابی داده و اطلاعاتی است که با استفاده از ابزارهای تحلیلی روی آمارهای به‌دست‌آمده انجام می‌شود و کشف این اطلاعات در گرفتن تصمیم‌های درست و پیش‌بینی‌ها در مسائل مختلف تأثیرگذار خواهد بود. آنالیز بقا یکی از مجموعه‌های روش‌های آماری تحلیل داده است که در آن متغیر مورد ارزیابی زمان وقوع یک رویداد است. هدف از آنالیز بقا

[DOI: 10.61186/jsdp.20.3.3]

یافتن یک مدل مناسب برای ارتباط بین مدت‌زمان زنده‌ماندن بیمار و علائم بالینی و ویژگی‌های بیمار است تا با کنترل بیماری و درمان‌های مؤثر مدت‌زمان زنده‌ماندن بیمار افزایش یابد [۱۰, ۱۱].

برای بررسی آنالیز بقا از مدل‌های آماری که به سه دسته مدل‌های پارامتری و مدل‌های غیر پارامتری و مدل‌های شبه‌پارامتری تقسیم می‌شوند، استفاده می‌شود. در مدل‌های پارامتری از توزیع پارامتریک مانند توزیع exponential و توزیع weibull و توزیع log-logistic برای توزیع زمان بقا استفاده می‌شود [۱۲]. در مدل‌های غیر پارامتری مانند مدل kapalan-meier که هیچ پیش‌فرضی در مورد توزیع زمان بقا در نظر گرفته نمی‌شود، که در آن تابع بقا به‌وسیله تخمین‌گر kapaln-meier تخمین زده می‌شود [۱۳]. این مدل بسیار ساده و توانایی مشارکت ویژگی‌ها را در مدل نداشته و مدل غیر پارامتری the life table که تخمینی از مدل kapalan-meier است؛ و برای دادگان‌های بزرگ بزرگ مورداستفاده قرار می‌گیرد. پرکاربردترین مدل آماری مدل رگرسیونی شبه‌پارامتری cox proportional hazard(CPH) است که توسط کاکس ارائه‌شده است. در این مدل هیچ پیش‌فرضی در مورد توزیع زمان بقا در نظر گرفته نمی‌شود؛ ولی بر اساس فرض متناسب‌بودن خطر، ساخته‌شده است و از احتمالات جزئی برای تخمین پارامترها استفاده می‌شود. در مدل‌های آماری فرض بر این است که ویژگی‌ها رابطه خطی با یکدیگر دارند [۱۴]. با توجه به اینکه در بسیاری از مطالعات و در واقعیت داده‌ها ممکن است، دارای ساختار پیچیده و تعاملات غیرخطی با ابعاد بالا باشند؛ بنابراین نیاز به مدل‌های غیرخطی پیچیده‌تری برای تخمین تابع خطر است. که با به‌کارگیری دیدگاه‌های یادگیری ماشین که اغلب الگوریتمی هستند و تعاملات بین ویژگی‌ها با آموزش استخراج می‌شوند، این نیاز برآورده شده است. مدل جنگل تصادفی بقا برای انتخاب متغیرهای مهم و تأثیرگذار بر بقا معرفی شد که می‌تواند در برابر این مشکلات مدل‌های آماری توانمند باشد [۱۵]. دو برتری عمدهٔ این روش نسبت به CPH ناپارامتری بودن کامل آن و توانایی سنجش تمام اثرات یک و چند متغیر به‌طور خودکار است و همچنین این روش قادر است تا ویژگی‌های مؤثر بر بقا در یک مجموعه‌ای از ویژگی‌ها با همبستگی بالا را شناسایی کند. مدل‌های دیگری از یادگیری ماشین که برای بررسی آنالیز بقا به کار می‌روند بر اساس شبکه‌های عصبی مصنوعی هستند. که یک شبکه عصبی معمولی از یک لایه ورودی ، چند لایه مخفی و یک لایه خروجی تشکیل شده‌است که لایه خروجی، خروجی را رده‌بندی می‌کند یا مقادیر مختلفی رگرسیونی را پیش‌بینی می‌کند. شبکه‌های عصبی توانایی یادگیری تعاملات غیرخطی بین ویژگی‌ها را دارند و کارایی بالایی را در آنالیز بقا با حذف محدودیت‌های مدل‌های آماری از خود نشان دادند. مدل ابتدایی از کاربرد شبکه‌های عصبی در آنالیز بقا یک مدل بسیار ساده بدون هیچ لایه مخفی برای پیش‌بینی تابع خطر بود [۱۶]. فراگی و همکارانش یک شبکه عصبی مصنوعی ساده با یک گره خروجی و یک لایه پنهان ارائه دادند که گره خروجی تابع خطر(log-risk) در مدل کاکس را پیش‌بینی می‌کرد. در آن از احتمالات جزئی به‌عنوان تابع ضرر برای آموزش استفاده می‌کرد. در این مدل داده‌های سانسور شده از راست[1] مشارکت داده شدند [۱۷]. در مدل دیگری ارائه داد که شامل گره‌های خروجی چندگانه و فواصل زمانی بود و از زمان‌های بقای گروهی و تخمین‌گر kapaln-meier برای احتمال بقا و داده‌های سانسورشده از راست در آن استفاده کرد. و مدل به کمک تابع ضرر cross entropy آموزش داده شد، این مدل عملکرد خوبی را در مقایسه با مدل‌های قبلی داشته است [۱۸]. فوتسو (۲۰۱۸) برخلاف یو و همکارانش (۲۰۱۱) که تابع بقا با ترکیب چند مدل رگرسیون منطقی که هرکدام تابع بقا را برای فاصله زمانی خاصی پیش‌بینی می‌کنند، تخمین زده می‌شود و تعاملات غیرخطی بین ویژگی‌ها در نظر گرفته نمی‌شود [۱۹]، از شبکه‌های عصبی به‌عنوان هستهٔ مدل استفاده می‌شود و بدین ترتیب توانایی مدل را برای تعاملات غیرخطی بین ویژگی‌ها افزایش می‌یابد [۲۰]. کاتزمن و همکارانش (۲۰۱۸) دیدگاه Deepsurv ارائه کردند که در آن از شبکه‌های عصبی پیچیده‌تری به همراه روش‌های یادگیری عمیق استفاده می‌شود که قادر به یادگیری تعاملات غیرخطی بین ویژگی‌ها است [۲۱]. این مدل بر اساس معیار c_index concordance_index (معیاری برای سنجش قابلیت مدل‌های آنالیز بقا است [۲۲]) نسبت به مدل کاکس عملکرد بهتری را دارد. آنالیز بقا قابلیت وفق داشتن با داده‌های سانسورشده از سمت راست را دارد و معیار (c_index) توانایی اعمال این داده‌ها در بررسی احتمال پیش بینی صحیح مدل را دارد و برای سنجش قابلیت مدل‌های آنالیز بقا مورد بررسی قرار می‌گیرد. بسیاری از کارهای قبلی با مشاهده زمان بقا به‌عنوان نخستین زمان برخورد یک فرآیند تصادفی، با فرض یک

[1] right-censored

[DOI: 10.61186/jsdp.20.3.3]

فرم خاص برای فرآیند تصادفی اساسی، با استفاده از داده‌های موجود برای یادگیری رابطه بین متغیرهای کمکی و پارامترهای مدل و سپس استنتاج رابطه بین متغیرهای کمکی و توزیع زمان‌های برخورد اول (ریسک) به مسئله نزدیک شده‌اند. بااین‌حال، مدل‌های قبلی بر فرضیات پارامتری قوی تکیه می‌کنند که اغلب نقض می‌شوند. بنا به همین موارد لی و همکاران (۲۰۱۸) یک روش بسیار متفاوت برای تجزیه‌وتحلیل بقا، DeepHit، پیشنهاد می‌کنند که از یک شبکه عصبی عمیق برای یادگیری توزیع زمان‌های بقا به‌طور مستقیم استفاده می‌کند [۲۳]. DeepHit هیچ فرضی در مورد فرآیند تصادفی اساسی ایجاد نمی‌کند و این امکان را فراهم می‌کند که رابطه بین متغیرهای کمکی و ریسک‌ها در طول زمان تغییر کند.

این مطالعه در نظر دارد که میزان بقای بیماران مبتلا به سرطان پستان را با استفاده از نسخه بهینه‌سازی شده DeepHit انجام دهد تا دقت آن به میزان قابل‌توجهی افزایش یابد. در این کار با تغییراتی که در تعداد و اندازه‌های لایه‌های مربوط به مدل پایه و همچنین تغییراتی مربوط به ضرایب Loss function می‌شود سبب بهبود نتایج نسبت به کارهایی مشابهی شده است که تنها از همان مدل پایه استفاده کرده‌اند.

۲- مواد و روش‌ها

داده‌های مورداستفاده از بانک‌داده‌ای METABRIC مربوط به طبقه‌بندی مولکولی از بیماران مبتلا به سرطان سینه مجمع بین‌المللی است [۲۴]، که شامل پروفایل ژنی و ویژگی‌های کلینیکی بیماران است (شماره دسترسی: EGAS00000000083, http://www.ebi.ac.uk/ega). تعداد کل بیماران موردبررسی ۱۹۸۱نفر است. از این تعداد ۴۴.۸٪ (۸۸۸ نفر) از بیماران تا لحظه مرگ تحت مراقبت بوده‌اند و بقیه بیماران (۱۰۹۳ نفر) در حین مطالعه از ادامه مطالعه صرف‌نظر کرده‌اند و از آن‌ها به‌عنوان داده‌های سانسورشده از راست در مطالعه استفاده می‌شود. در این مجموعه‌دادگان به ویژگی‌های بالینی بیماران از قبیل اندازهٔ تومور، تعداد غدد لنفاوی مثبت و حذف‌شده، گیرنده پروژسترون، گیرنده استروژن، فاکتور رشد اپیدرمی، ویژگی مارکرهای ایمونوهیستوشیمی و پروفایل ژنی بیماران مانند سن تشخیص بیماری توجه شده است که در کل شامل ویژگی های کمی و کیفی است. ویژگی های کمی این پایگاه داده در جدول (۱) نشان داده شده است.

(جدول- ۱): مشخصات کمی داده‌های METABRIC مربوط به طبقه‌بندی مولکولی از بیماران مبتلابه سرطان سینه

(Table-1): Quantitative Characteristics of METABRIC data related to molecular classification of breast cancer patients.

داده	کمینه	بیشینه	میانگین	انحراف معیار
Age at diagnosis	۲۱	۹۲	۶۰/۶۴	۱۳/۱
size	۰	۱۸۲	۲۵/۳۹	۱۴/۳۶
Lymph nodes positive	۰	۴۵	۱/۸	۳/۸
Lymph nodes removed	۰	۴۸	۱۲/۸۲	۸/۲۴
NPI	۱	۶/۳	۳/۹۹	۱/۱۳
Event time	۳	۹۲۱۸	۲۹۵۱	۱۷۹۶/۳

ویژگی‌های کیفی به‌صورت ONE-HOT کد شده‌اند. داده‌های گمشده در این بانک داده برای ویژگی‌های کمی با میانگین دیتاهای واقعی جایگزین شده‌اند. همچنین زمان وقوع مرگ و زمان سانسور در این بانک داده برای بیماران ثبت شده‌است. ویژگی های کیفی پایگاه داده مورد نظر در جدول (۲) نشان داده شده است.

برای استفاده از این بانک داده ابتدا لازم است تغییراتی بر روی آن اعمال شود. عملیات نرمال‌سازی بر روی‌داده‌های مربوط به ویژگی‌ها انجام شده و داده‌هایی که همبستگی آن‌ها بیشتر از ۰/۹۵ است، حذف شده‌اند. همچنین برای عملکرد بهتر مدل بر روی داده‌های مربوط به ویژگی‌ها تبدیل KPCA اعمال شده و ابعاد ویژگی‌ها به پانزده عدد کاهش‌یافته است. تحلیل مؤلفه اساسی به روش کرنل، بر بسیاری از محدودیت‌های روش خطی PCA به‌وسیله نگاشت غیرخطی فضای ورودی به یک فضای ویژگی با ابعاد بالا غلبه می‌کند. خطی‌بودن در فضای ویژگی ولی غیرخطی‌بودن آن در فضای ورودی، KPCA را قادر به استخراج ویژگی‌های با ابعاد پایینی که در اطلاعات آماری مرتبه بالا وجود دارند، می‌کند [۲۵].

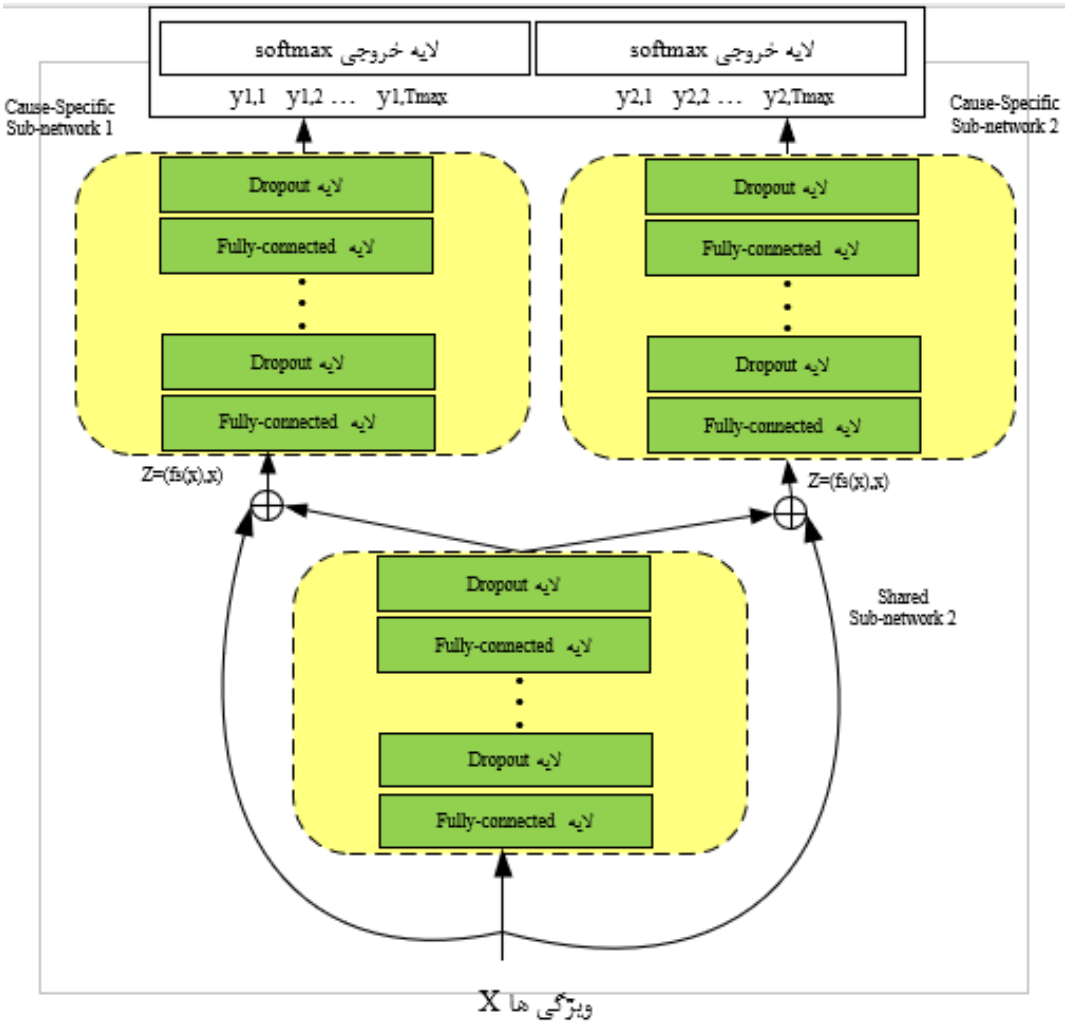
مدل DeepHit، یک شبکه عصبی یادگیری عمیق است که بر اساس الگوریتم back propagation عمل می‌کند و بر مبنای یادگیری multi-task طراحی شده‌است و هیچ پیش‌فرضی در آن برای مدل و داده‌ها در نظر گرفته نمی‌شود. همان‌طور که در شکل (۱) معماری مدل نشان داده شده‌است، شبکه شامل دو بخش است: بخش نخست مربوط به زیرشبکه اشتراکی است که از یک لایه تمام متصل (همان‌طوری که از نام این لایه پیداست تمامی نرون‌ها این لایه به نرون‌های لایهٔ قبل متصل هستند. وظیفه اصلی آن ترکیب ویژگی محلی در لایه پایین به ویژگی محلی در لایه‌های بالاست) که به‌وسیله یک لایه بیرون‌انداز drop out (به معنای کنار گذاشتن بخش‌هایی (units) از یک شبکه عصبی است. یعنی در حین آموزشِ نورون‌ها، از تعدادی از آن‌ها به صورت تصادفی چشم‌پوشی شود.

[DOI: 10.61186/jsdp.20.3.3]

(جدول- ۲): مشخصات کیفی داده‌های METABRIC مربوط به طبقه‌بندی مولکولی از بیماران مبتلابه سرطان سینه

(Table-2): Qualitative Characteristics of METABRIC data related to molecular classification of breast cancer patients.

	نحوه توزیع : مقدار(تعداد)					
Expr Her2	۰(۲۳۴)	۱(۱۷۴۷)				
ER Expr	۰(۱۵۴۵)	۱(۴۳۶)				
Expz PR	۰(۱۰۹۱)	۱(۸۹۰)				
status IHC ER	۰(۴۰۴)	۱(۱۵۷۷)				
status men Inf	۰(۱۵۰۲)	۱(۴۷۹)				
cellularity	۰(۱۰۵۲)	۱(۲۱۶)	۲(۷۱۳)			
grade	۰(۱۷۰)	۱(۷۶۷)	۲(۱۰۴۴)			
status IHC HER2	۱(۱۸۴۰)	۱(۳۰)	۳(۱۱۱)			
HER2 SNP6 state	۰(۴۱۸)	۱(۹۷)	۲(۱۴۶۶)			
Genefu	۰(۱۳۹۵)	۱(۳۶۳)	۲(۱۲۷)	۳(۹۶)		
Treatment	۰(۴۸) ۱(۳۱)	۲(۱۵۲) ۳(۱۶۰)	۴(۴۰۷) ۵(۶۶۰)	۶(۲۹۹) ۷(۲۲۴)		
group	۰(۷۷۳)	۱(۴۳۳)	۲(۲۳۷)	۳(۵۳۱)	۴(۷)	
stage	۰(۴۶۴)	۱.۱۹ (۸۶۸)	۲(۵۵۸)	۳(۸۳)	۴(۱۰)	
site	۰(۵۷۰)	۱(۲۷۰)	۲(۷۳۳)	۳(۲۳۷)	۴(۱۷۱)	
Int clust memb	۰(۶۹) ۱(۴۲)	۲(۱۵۲) ۳(۱۲۱۲)	۴(۸۲) ۵(۴۲)	۶(۱۰۵) ۷(۱۳۹)	۸(۶۱) ۹(۷۷)	
Pam50 Subtype	۰(۳۰۳)	۱(۲۲۸)	۲(۷۷۶)	۳(۴۷۰)	۴(۶)	۵(۱۹۸)
histological	۰(۱۰) ۱(۱۵۵۹)	۲(۸۹) ۳(۲۸)	۴(۴۶) ۵(۶۵)	۶(۱۴۴) ۷(۹)	۸(۲) ۹(۱۲)	۱۰(۱۲) ۱۱(۵)



(شکل- ۱): معماری الگوریتم شبکه عصبی عمیق DeepHit

(Figure-1): DeepHit deep neural network algorithm architecture

[DOI: 10.61186/jsdp.20.3.3]

چشم‌پوشی یعنی اینکه آن نورون‌های خاص، در مسیر رفت یا برگشت در نظر گرفته نمی‌شوند) دنبال می‌شود، تشکیل شده‌است. بخش دوم مربوط به گروهی از زیرشبکه‌های دلایل خاص است که به‌ازای هر رویداد شامل یک لایه تمام متصل است که با یک لایه بیرون انداز drop out دنبال می‌شود. به دلیل داشتن این ساختار، این مدل به‌راحتی می‌تواند برای دادگان‌هایی با یک ریسک یا چند ریسک رقابتی مورداستفاده قرار بگیرد. ساختار و معماری مدل بستگی به تعداد ریسک‌ها یا رویدادهای رقابتی دارد. در این مدل یک ارتباط باقیمانده بین ویژگی‌ها (ورودی اصلی) و ورودی شبکه‌های دلایل خاص وجود دارد. یعنی ورودی زیر شبکه‌های دلایل خاص علاوه‌بر خروجی زیرشبکه اشتراکی شامل ورودی اصلی هم است. این ورودی اضافی به زیرشبکه‌های دلایل خاص اجازه می‌دهد که ویژگی‌های غیرمشترک از دلایل چندگانه را بهتر یاد بگیرند. در خروجی مدل از یک لایه softmax استفاده می‌شود که توزیع توأم از رویدادهای رقابتی را به‌جای توزیع حاشیه‌ای یاد می‌گیرد؛ بنابراین خروجی مدل یک بردار y برای هر نمونه در دادگان با ویژگی x است که احتمال تجربه رویداد k را در زمان t نمایش می‌دهد [۲۳].

هر زیرشبکه علت خاص[1] به‌عنوان ورودی جفت $z=(f_s(x),x)$ در نظر گرفته می‌شود و به‌عنوان خروجی یک بردار $f_{ck}(z)$ تولید می‌کند، که مربوط به احتمال نخستین زمان برخورد یک علت خاص k است. به‌طور خاص‌تر، ورودی‌ها به زیرشبکه‌ها شامل هر دو خروجی شبکه مشترک و متغیرهای کمکی اصلی هستند؛ این امر به زیرشبکه‌ها امکان دسترسی به بازنمود مشترک یاد گرفته‌شده $f_s(x)$ را می‌دهد و درعین‌حال امکان یادگیری بخش غیرمعمول بازنمود را نیز به آن‌ها می‌دهد. اگر فقط نمایش مشترک آموخته‌شده به‌عنوان ورودی به زیرشبکه‌ها استفاده شود، بخش غیرمعمول نمایش از دست خواهد رفت. کلیت این خروجی‌ها یک توزیع احتمال مشترک در نخستین زمان برخورد و رویداد است؛ بنابراین زیرشبکه‌های علت خاص توزیع را برای نخستین‌بار برای هر علت به‌صورت موازی یاد می‌گیرند. خروجی لایه softmaxتوزیع احتمال y است. یک بیمار با متغیرهای کمکی x، یک عنصر خروجی $y_{k,s}$ احتمال $\hat{P}(s,k|x)$ است که بیمار رویداد k را در زمان s تجربه خواهد کرد؛ بنابراین این معماری، شبکه را به سمت یادگیری روابط غیرخطی و حتی غیرمتناسب بین متغیرهای کمکی و ریسک‌ها سوق می‌دهد [۲۳].

در این مدل دو تابع ضرر loss log likelihood و loss Ranking برای آموزش داریم که طبق رابطه (۱) محاسبه می‌شود:

$$L_{total} = L_1 + L_2 \quad (۱)$$

که در آن $\mathcal{L}_1$ تابع شبه‌لگاریتم یا loss log likelihood از توزیع توأم نخستین زمان وقوع رویداد است این تابع به‌گونه‌ای اصلاح شده‌است که داده‌های سانسورشده و ریسک‌های رقابتی را در برمی‌گیرد. تابع شبه‌لگاریتم شامل دو ترم است. بخش نخست، هم رویداد و هم‌زمان وقوع را برای داده‌های سانسورنشده و در بخش دوم زمان سانسور را برای داده‌های سانسورشده در نظر می‌گیرد، که از طریق رابطه (۲) محاسبه می‌شود:

$$L_1 = -\sum_{i=1}^{k}[F\left(k^{(i)} \neq \varphi\right).\log(y_{k(i),s(i)}^{(i)}) + F(k^{(i)} = \varphi).\log(1-\sum_{k=1}^{k}\hat{F}_k(s^{(i)}|X^{(i)}))] \quad (۲)$$

که در آن $\mathbb{F}(.)$ یک تابع شاخص است [۲۳].

همچنین $\mathcal{L}_2$ ترکیبی از تابع‌های ضرر درجه‌بندی دلایل خاص یا loss Ranking است. به‌دلیل اینکه این مدل یک مدل یادگیری multi-task است، نیاز به تابع‌های ضرر دلایل خاص دارد که این تابع همان تابع رویدادهای تجمعی یعنی CIF است. این تابع احتمال اینکه رویداد خاص k قبل یا به‌طوردقیق در زمان t به همراه ویژگی x رخ دهد را بیان می‌کند. برای تخمین CIF، مجموع احتمال‌ها از زمان مشاهده نخست تا زمان وقوع رویداد K محاسبه می‌شود، که از طریق رابطه (۳) محاسبه می‌شود:

$$L_2 = \sum_{k=1}^{K}\alpha_k \cdot \sum_{i\neq j} A_{k,i,j} \cdot \eta(\hat{F}_k\left(s^{(i)}|X^{(i)}\right), \hat{F}_k\left(s^{(i)}|X^{(j)}\right)) \quad (۳)$$

که در آن ضرایب α_k برای تنظیم و تعادل تلفات رتبه‌بندی رویداد رقابتی k ام انتخاب می‌شوند و $\eta(x,y)$ یک تابع اتلاف محدب است. گفتنی است که برای راحتی، در اینجا فرض می‌کنیم که ضرایب α همگی برابر هستند (یعنی $\alpha_k=\alpha$ برای k=1,…,K)، و از تابع اتلاف $\eta(x,y)=\exp(-(x-y)/\sigma)$ استفاده می‌کنیم. همچنین در آن $A_{k,i,j}$ یک تابع کاهش رتبه است که با ایده هماهنگی سازگار است: بیماری که در زمان s می‌میرد باید در زمان s ریسک بالاتری نسبت به بیماری داشته باشد که بیش از s جان سالم به در برده‌است. که به‌صورت زیر تعریف می‌شود:

[1] cause-specific sub-network

[DOI: 10.61186/jsdp.20.3.3]

$$A_{(k,i,j)} \square \mathrm{F}(k^{(i)} = k, s^{(i)} < s^{(j)}) \quad (۴)$$

برای ارزیابی این مدل از معیار c_index استفاده می‌شود. این معیار توانایی مدل را برای فراهم‌آوردن یک درجه‌بندی قابل‌اعتماد از زمان‌های بقا بر اساس امتیازهای ریسکی نمونه‌ها نمایش می‌دهد. در این ارزیابی، مدل کل جفت‌های موجود و سپس جفت‌های concordant یعنی جفت‌های که زمان واقعی رویداد آن‌ها و زمان پیش‌بینی‌شده را توسط مدل برای آن‌ها یکسان باشد، شناسایی می‌کند. با تقسیم جفت‌های concordant به‌کل جفت‌های موجود، احتمال پیش‌بینی صحیح مدل را بررسی می‌کند [۲۳].

برای پیاده‌سازی این مدل از دو لایه تمام متصل با تابع فعالیت ساز relu و یک لایه بیرون‌انداز به‌عنوان لایه اشتراکی و همچنین از یک لایه تمام متصل با تابع فعالیت ساز relu و یک لایه بیرون‌انداز به‌عنوان لایه دلایل خاص استفاده شده‌است. و لایه خروجی از یک لایه تمام متصل با تابع فعالیت‌ساز soft max تشکیل شده‌است. تابع بهینه‌سازی Adam و تابع ضرر $\mathcal{L}_{total}$ برای آموزش به‌کاربرده شده‌است.

۳-یافته‌ها

برنامه‌نویسی مدل با استفاده از کتابخانه تنسورفلو۲ و در محیط گوگل‌کولب انجام شده‌است. جهت اجرای الگوریتم، داده‌ها به دو دسته، داده‌های آموزش و داده‌های آزمون با نسبت ۸۰ به ۲۰ تقسیم‌بندی شده‌اند. پارامترهای تنظیمی در جدول (۳) آورده شده اند.

با توجه به اینکه برای جلوگیری از بیش‌برازش از روش توقف زودهنگام استفاده شده‌است، که سبب شده فرایند یادگیری تا epoch=575 متوقف شود. نمودار مربوط به loss در شکل (۲) نشان داده‌شده است. گفتنی است درصورتی‌که میانگین نمودار loss کشیده شود، کاهش خطا به صورت یکنواخت خواهد بود.

همان‌طور که اشاره شد؛ به‌عنوان معیار عملکرد، از شاخص هماهنگی وابسته به زمان (C^{td}-index) استفاده شد [۲۲]. به یاد داشته باشید که شاخص هماهنگی معمولی (C-index)یک شاخص متمایزکننده است که به‌طور گسترده‌ای بر اساس این فرض است که بیمارانی که عمر طولانی‌تری دارند، باید نسبت به بیمارانی که عمر کمتری دارند، ریسک کمتری داشته باشند.C^{td}-index برای رویداد k به‌صورت زیر تعریف می‌شود:

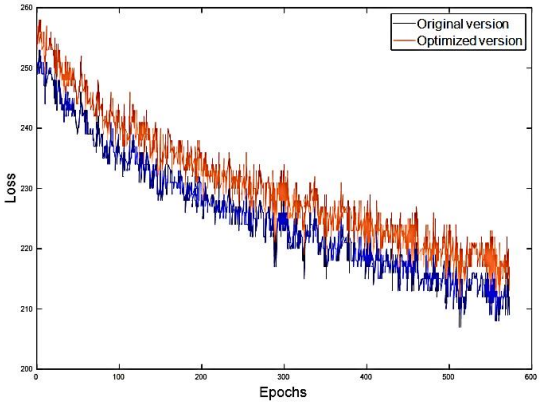
$$Ctd = p(\hat{F}_k(s^{(i)} \mid X^{(i)}) > \hat{F}_k(s^{(i)} \mid X^{(j)}) \mid s^{(i)} \mid s^{(j)}) \approx \frac{\sum_{i \neq j} A_{k,i,j} . \mathrm{F}\left(\hat{F}_k(s^{(i)} \mid X^{(i)}) > \hat{F}_k(s^{(i)} \mid X^{(j)})\right)}{\sum_{i \neq j} A_{k,i,j}} \quad (۵)$$

که در آن، $A_{k,i,j}$ تابع شاخص برای یک جفت (i,j) است که برای یک رویداد k قابل‌قبول است و تقریب از تعریف تجربی حاصل می‌شود؛ بنابراین، شاخص C^{td} برای رویداد k از مقایسه جفت‌هایی که در آن یک بیمار رویداد k را در یک‌زمان خاص تجربه کرده است به‌دست می‌آید، درحالی‌که دیگری هیچ رویدادی را تجربه نکرده و در آن زمان، سانسور نشده است. نتایج مربوط به شاخص C^{td} در جدول ۴ نمایش داده‌شده است که با شبکه عصبی DeepHit در کارهای گذشته و دیگر یادگیری‌های ماشین مقایسه شده است.

(جدول-۳) : مشخصات پارامترهای تنظیم‌شده جهت اجرای شبکه عصبی عمیق Deephit بهینه‌شده

(Table-3): Specifications of tuned parameters to run the optimized Deep Hit deep neural network

پارامترها	Size Batch	Learning Rate	epochs	Number of Foley connection layers
مقدار	1024	0.0001	1000	60



(شکل-۲): مقدار تابع خطا در فرایند یادگیری

(Figure-2): The value of the loss function in the learning process

۴- بحث

سرطان پستان یک مسئله مهم اپیدمیوژیک با گسترش جهانی و یکی از مهم‌ترین علل مرگ‌ومیر و درواقع دومین علت مرگ براثر سرطان‌ها در زنان است. استفاده از روش‌های نوین مبتنی بر هوش مصنوعی به علت توانایی در کاهش پیچیدگی روابط مدل‌ها و قابلیت افزایش یادگیری مدل با تشخیص بهتر و دقت بالا در پیش‌بینی در مطالعات و پژوهش‌های اخیر افزایش قابل‌ملاحظه‌ای دارد.

[DOI: 10.61186/jsdp.20.3.3]

(جدول -۴): مقایسه نتایج مربوط به شاخص C^{td} در مدل ارائه بهینه‌سازی شده

(Table-4): Comparison of the results corresponds to the C^{td}-index in the proposed optimized model

الگوریتم‌ها	Cox	RSF	ThresReg	MP-RForest	MP-AdaBoost	MP-LogitR	DeepSurv	DeepHit [۲۳]	Deephit بهینه‌سازی شده (روش پیشنهادی)
نتایج C^{td}-index	0.648	0.672	0.649	0.650	0.633	0.661	0.648	0.691	**0.73**

در این مطالعه، مدل شبکه عصبی عمیق DeepHit بهینه‌سازی شده ارائه شد. همان‌طوری‌که از نتایج مطالعات گذشته مشخص می‌شود، میزان بقا به‌عنوان یکی از مهم‌ترین شاخص‌ها است که با برآوردی از پیش‌آگهی بیماری، در ارائه یک روش تشخیصی و درمانی مناسب به کمک جامعه پزشکی می‌آیند.

داده‌کاوی قادر به کشف و استخراج دانش جدید از داده‌های گذشته‌نگر است نحوه‌ی پیش‌پردازش داده‌ها و همچنین متغیرهای منتخب تأثیر قابل‌توجهی در کشف دانش دارد که در این مدل از KPCA استفاده شد. مدل بهینه‌سازی‌شده دارای تغییراتی از قبیل افزایش تعداد لایه‌های مربوط به تمام متصل به‌اندازه ۶۰، تغییر در batch size به مقدار ۱۰۲۴، استفاده از مدل KPCA، و تنظیم پارامتر بهینه α است که سبب افزایش چشم‌گیری در نتایج C^{td} نسبت به کارهای گذشته شده است. ازآنجاکه شاخص متمایزکننده C^{td} به یک‌زمان ثابت واحد بستگی ندارد، یک ارزیابی مناسب برای شرایطی فراهم می‌کند که در آن تأثیر متغیرهای کمکی بر بقا در طول زمان تغییر می‌کند (به‌عبارت‌دیگر، ریسک‌ها در طول زمان غیرمتناسب هستند) [۲۲].

برای مجموعه‌داده‌های METABRIC که دارای یک رویداد واحد (ریسک) هستند، عملکرد متمایز DeepHit بهینه‌سازی‌شده با دو خانواده از دیگر مدل‌های بقا مقایسه شد. خانواده‌ای از مدل‌های بقا تشکیل‌شده‌اند که از پیش‌بینی مرگ‌ومیر اجراشده توسط الگوریتم‌های یادگیری ماشین مشتق شده‌اند: جنگل‌های تصمیم تصادفی (MP-RFoest)، رگرسیون ترابری (MP-LogitR)، و الگوریتم تقویت‌کننده (MP-AdaBoost) و با شبکه عصبی عمیق (Deepsurv)، که بر اساس فرض متناسب کاکس توسعه داده شده‌است [۲۶]. همان‌طور که دیده شد، DeepHit بهبودیافته، عملکرد را نسبت به مدل‌های دیگر فراهم می‌کند.

۵- نتیجه‌گیری

آنالیز بقا ابزار مفید و کارآمدی در پزشکی برای تحقیقات کلینیکی است و اطلاعات با ارزشی را در مورد ارزیابی انواع درمان‌ها و کنترل بیماری و پیش‌بینی‌ها و غیره فراهم می‌کند. شبکه‌های عصبی مصنوعی با موفقیت برای تشخیص الگو و پیش‌بینی درزمینه‌های بالینی استفاده‌شده‌اند. یکی از مزایای استفاده از مدل شبکه عصبی، بررسی ارتباطات غیرخطی و اثرات متقابل پیچیده بین عوامل است. این مقاله یک روش جدید به نام DeepHit بهینه‌سازی شده را برای تجزیه‌وتحلیل داده‌های بقا ارائه می‌دهد. DeepHit یک شبکه عصبی را آموزش می‌دهد تا توزیع مشترک برآورد شده زمان بقا و رویداد را یاد بگیرد، درحالی‌که طبیعت سانسور شده درست ذاتی در داده‌های بقا را ثبت می‌کند. ما شبکه را با استفاده از یک تابع زیان که از زمان بقا و ریسک‌های نسبی بهره می‌برد، آموزش می‌دهیم. به‌عنوان یک آزمون، ما عملکرد DeepHit بهینه‌سازی شده را با عملکرد مدل‌های قبلی مقایسه کردیم. در تنظیمات با ریسک‌های رقابتی، عملکرد DeepHit بسیار بهتر از مدل‌های قبلی است. برای کارهای آینده میتوان از الگوریتم های بهینه سازی برای بدست آوردن متغیرهای مورد نظر در تابع هزینه استفاده کرد.

۶- مراجع 6-Refrence

[1]. El-Bendary, N. and N.A. Belal, A feature-fusion framework of clinical, genomics, and histopathological data for METABRIC breast cancer subtype classification. Applied Soft Computing, vol.91, pp.106238, 2020.

[2]. Singh, D. and A.K. Singh, Role of image thermography in early breast cancer detection-Past, present and future. Computer methods and programs in biomedicine, vol. 183, pp.105074, 2020.

[3]. Azamjah N, Soltan-Zadeh Y, Zayeri F. Global trend of breast cancer mortality rate: a 25-year study. Asian Pacific journal of cancer prevention: APJCP. 2019;20(7):2015.

[4]. Forouzanfar, M.H., et al., Breast and cervical cancer in 187 countries between 1980 and 2010: a systematic analysis. The lancet, vol. 378(9801), pp.1461-1484, 2011.

[5]. Delen, D., G. Walker, and A. Kadam, Predicting breast cancer survivability: a comparison of three data mining methods. Artificial intelligence in medicine, vol.34(2), pp.113-127, 2005.

[6]. Campone, M., et al., Taxanes in adjuvant breast

[DOI: 10.61186/jsdp.20.3.3]

cancer setting: which standard in Europe? Critical reviews in oncology/hematology, vol.55(3), pp.167-175, 2005.
[7]. Mohammadpour, A., et al., Breast Cancer, Genetic Factors and Methods of Diagnosis. Sarem Journal of Reproductive Medicine,vol. 4(4): 198-207, 2020.
[8]. Mukherjee, A., et al., Associations between genomic stratification of breast cancer and centrally reviewed tumour pathology in the METABRIC cohort. NPJ breast cancer, vol.4(1) ,pp. 1-9, 2018.
[9]. Rakha, E.A. and A.R. Green, Molecular classification of breast cancer: what the pathologist needs to know. Pathology, vol.49(2), pp.119, 2017.
[10]. Hao, J., et al., Interpretable deep neural network for cancer survival analysis by integrating genomic and clinical data. BMC medical genomics,vol. 12(10) ,pp. 1-13,2019.
[11]. Collett, D., Modelling survival data in medical research. 2015: CRC press.
[12]. Stevenson, M. and I. EpiCentre, An introduction to survival analysis. EpiCentre, IVABS, Massey University,2009.
[13]. Goel, M.K., P. Khanna, and J. Kishore, Understanding survival analysis: Kaplan-Meier estimate. International journal of Ayurveda research,vol.1(4), pp.274, 2010.
[14]. Therneau, T.M. and P.M. Grambsch, The cox model, in Modeling survival data: extending the Cox model,pp. 39-7,2000.
[15]. O'Brien, R.C., et al., Random Survival Forests Analysis of Intraoperative Complications as Predictors of Descemet Stripping Automated Endothelial Keratoplasty Graft Failure in the Cornea Preservation Time Study. JAMA ophthalmology,vol.139(2) ,pp.191-197,2021.
[16]. Gensheimer MF, Narasimhan B. A scalable discrete-time survival model for neural networks. PeerJ. 2019 Jan 25;7:e6257.
[17]. Faraggi, D. and R. Simon, A neural network model for survival data. Statistics in medicine,vol.14(1) ,pp.73-82, 1995.
[18]. Street, W.N. A Neural Network Model for Prognostic Prediction. in ICML. 1998. Citeseer.
[19]. Fotso, S., Deep neural networks for survival analysis based on a multi-task framework. arXiv preprint arXiv:1801.05512, 2018.
[20]. Yu, C.-N., et al., Learning patient-specific cancer survival distributions as a sequence of dependent regressors. Advances in neural information processing systems, vol.24, pp.1845-1853, 2011.
[21]. Katzman, J.L., et al., DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC medical research methodology, vol.18(1), pp.121, 2018.
[22]. Luck M, Sylvain T, Cardinal H, Lodi A, Bengio Y. Deep learning for patient-specific kidney graft survival analysis. arXiv preprint arXiv:1705.10245. 2017 May 29.
[23]. Lee, C., et al. Deephit: A deep learning approach to survival analysis with competing risks. in Thirty-second AAAI conference on artificial intelligence,2018.
[24]. Bilal, E., et al., Improving breast cancer survival analysis through competition-based multidimensional modeling. PLoS computational biology, vol.9(5), pp.10030-47, 2013.
[25]. Lu, J., et al., An efficient kernel discriminant analysis method. Pattern Recognition, vol. 38(10) ,pp.1788-179,2005.
[26]. Lee, M.-L.T. and G.A. Whitmore, Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. Statistical Science,21(4) ,pp.501-513, 2006.

سهیلا رضایی، مدرک کارشناسی خود را در رشته برق-الکترونیک در سال ۱۳۹۰ از دانشگاه ملی هرمزگان و در سال ۱۴۰۱ مدرک کارشناسی ارشد را از دانشگاه ولی عصر(عج) رفسنجان دریافت کرده است. زمینه پژوهشی مورد علاقه ایشان عبارتند از: آنالیز داده، یادگیری ماشین، یادگیری عمیق و آنالیز بقا.
نشانی رایانامه ایشان عبارت است از:
s.rezaie2261368@gmail.com

حسین قیومی زاده، مدرک کارشناسی خود را در رشته مهندسی برق الکترونیک در سال ۱۳۸۸ از دانشگاه شهید رجائی تهران و مدارک کارشناسی ارشد و دکترای خود را در رشته مهندسی پزشکی (بیوالکتریک) به‌ترتیب در سال‌های ۱۳۹۰ و ۱۳۹۵ از دانشگاه حکیم سبزواری دریافت کرده است. وی در حال حاضر دانشیار دانشگاه ولی عصر(عج) در گروه مهندسی برق است. زمینه های پژوهشی ایشان عبارتند از: هوش مصنوعی، شبکه‌های عصبی عمیق و پردازش تصویر.
نشانی رایانامه ایشان عبارت است از:
h.ghayoumizadeh@vru.ac.ir

[DOI: 10.61186/jsdp.20.3.3]

محمدحسین قلی‌زاده، مدرک کارشناسی خود را در رشته مهندسی برق مخابرات در سال ۱۳۸۶ از دانشگاه صنعتی اصفهان و مدارک کارشناسی ارشد و دکترای خود را در رشته مهندسی برق مخابرات (سیستم) به ترتیب در سال‌های ۱۳۸۸ و ۱۳۹۴ از دانشگاه صنعتی امیرکبیر دریافت کرده است. زمینه‌های پژوهشی ایشان عبارتند از: هوش مصنوعی، شبکه‌های عصبی و پردازش تصویر.
نشانی رایانامه ایشان عبارت است از:
gholizadeh@vru.ac.ir

علی فیاضی، مدرک کارشناسی خود را در رشته مهندسی برق-الکترونیک در سال ۱۳۸۵ برق از دانشگاه شهید باهنر کرمان و در سال ۱۳۸۷ مدرک کارشناسی ارشد خود را در رشته مهندسی برق-کنترل از دانشگاه آزاد واحد علوم و تحقیقات تهران و در سال ۱۳۹۷ مدرک دکترای تخصصی خود را در رشته مهندسی برق-کنترل از دانشگاه فردوسی مشهد دریافت کرده است. وی مدرس دانشگاه ولی عصر(عج) دانشگاه رفسنجان از سال ۱۳۸۸ تا ۱۳۹۱ بود و از سال۱۳۹۶ به عنوان استادیار گروه مهندسی برق دانشگاه ولی عصر رفسنجان منصوب شد. زمینه‌های پژوهشی مورد علاقه ایشان عبارتند از: رباتیک، مکاترونیک، کنترل امپدانس، کنترل غیرخطی، کنترل تطبیقی، کنترل کسری و پردازش تصویر است.
نشانی رایانامه ایشان عبارت است از:
a.fayazi@vru.ac.ir

[DOI: 10.61186/jsdp.20.3.3]