Published by: Iranian Society for Biomedical Engineering / www.isbme.ir

Iranian Journal of Biomedical Engineering

www.ijbme.org / P-ISSN: 2008-5869 / E-ISSN: 8006-9685

Volume 19, Issue 2, Summer 2025, 151 - 164

Detection of Epidermal Growth Factor Receptor Mutations in Non-Small Cell Lung Cancer Patients Using a Supervised Representation Learning Framework

Bahrami, Mahsa [1] / Vali, Mansour [2*] / Kazemizadeh, Hossein[3]

[1] - Ph. D. Student, Department of Biomedical Engineering, Faculty of Electrical Engineering, K. N. Toosi University of Technology, Tehran, Iran
[2] - Associate Professor, Department of Biomedical Engineering, Faculty of Electrical Engineering, K. N. Toosi University of Technology, Tehran, Iran
[3] – Assistant Professor, Department of Pulmonology, Imam Khomeini Hospital Complex (IKHC), Tehran University of Medical Sciences, Tehran, Iran

ARTICLE INFO

DOI: https://doi.org/10.22041/ijbme.2025.2070816.1995
Received: 9-9-2025 Revised: 19-11-2025 Accepted: 30-11-2025

KEYWORDS

EGFR
Representation Learning
WXGB
Imbalanced Learning
Heterogeneous Data
NSCLC

ABSTRACT

Lung cancer remains one of the most prevalent malignancies worldwide and is a leading cause of cancer-related mortality. Accurate, automated detection of genetic mutations—particularly in the epidermal growth factor receptor (EGFR)—is essential for selecting targeted therapies and improving clinical outcomes in patients with non–small cell lung cancer (NSCLC). In recent years, machine learning methods have shown considerable promise in analyzing clinical data to identify genetic alterations. However, data heterogeneity and class imbalance in clinical datasets remain persistent challenges, leading to reduced predictive performance and biased models. In this study, we introduce a novel supervised representation learning framework specifically designed for heterogeneous clinical data comprising both categorical and numerical features. In this framework, categorical features are first encoded through a trainable embedding layer, while numerical data are preprocessed using a normalization layer. The learned embeddings are then integrated with preprocessed numerical features, and the combined inputs are passed through a fully connected layer to produce robust representations that capture complex relationships across heterogeneous data types. Finally, to address the class imbalance problem and improve the accuracy of minority class detection, a weighted XGBoost classifier is employed, which assigns different weights to classes to facilitate the identification of rare mutations. We evaluated the effectiveness of this framework on the NSCLC Radiogenomics dataset from The Cancer Imaging Archive (TCIA), which contains data from 211 patients. Five-fold Stratified cross-validation was employed to ensure model reliability. The proposed method achieved 80.9% accuracy, 72.2% sensitivity, 62.6% F1-score, 83.7% specificity, 66.5% precision, and an area under the ROC curve (AUC) of 0.82. Comparison with state-of-the-art methods demonstrated that the proposed method significantly improves EGFR mutation detection in heterogeneous and imbalanced clinical data.

[*]Corresponding Author

Address: Department of Biomedical Engineering, K. N. Toosi University of Technology, Tehran, Iran
E-Mail: mansour.vali@eetd.kntu.ac.ir

ناشر: انجمن مهندسی پزشکی ایران / www.isbme.ir

نشریه‌ی مهندسی پزشکی زیستی

شاپای چاپی: ۵۸۶۹-۲۰۰۸ / شاپای الکترونیکی: ۹۶۸۵-۸۰۰۶ / www.ijbme.org

دوره: ۱۹، شماره: ۲، تابستان ۱۴۰۴، ۱۵۱ - ۱۶۴

تشخیص جهش گیرنده رشد اپیدرمی در بیماران مبتلا به سرطان ریه سلول‌های غیرکوچک با استفاده از یک چارچوب یادگیری بازنمایی نظارت شده

بهرامی، مهسا [۱] / ولی، منصور [۲*] / کاظمی زاده، حسین [۳]

[۱] – دانشجوی دکتری، گروه مهندسی پزشکی، دانشکده مهندسی برق، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران

[۲] – دانشیار، گروه مهندسی پزشکی، دانشکده مهندسی برق، دانشگاه صنعتی خواجه نصیرالدین طوسی، تهران، ایران

[۳] – استادیار، گروه ریه، مجتمع بیمارستانی امام خمینی (ره)، دانشگاه علوم پزشکی تهران، تهران، ایران

مشخصات مقاله

شناسه‌ی دیجیتال: https://doi.org/10.22041/ijbme.2025.2070816.1995

ثبت در سامانه: ۱۴۰۴/۶/۱۸ بازنگری: ۱۴۰۴/۸/۲۸ پذیرش: ۱۴۰۴/۹/۹

واژه‌های کلیدی

گیرنده فاکتور رشد اپیدرمی
یادگیری بازنمایی
طبقه بند *WXGB*
یادگیری از داده‌های نامتوازن
دادگان ناهمگون
سرطان ریه سلول‌های غیرکوچک

چکیده

سرطان ریه یکی از شایع‌ترین سرطان‌های شناخته شده در جهان است که از مهم‌ترین علل مرگ و میر ناشی از سرطان محسوب می‌شود. شناسایی دقیق و خودکار جهش‌های ژنتیکی، به‌ویژه در گیرنده‌ی فاکتور رشد اپیدرمی***(EGFR)*** ، برای انتخاب درمان‌های هدفمند و بهبود شرایط بالینی در بیماران مبتلا به سرطان ریه سلول‌های غیر کوچک ***(NSCLC)*** ضروری است. در سال‌های اخیر، شناسایی جهش‌های ژنتیکی با استفاده از روش‌های یادگیری ماشین به افق‌های روشنی دست یافته است. با این حال، ناهمگونی داده‌ها و عدم توازن کلاس‌ها از چالش‌های مهمی هستند که منجر به کاهش عملکرد مدل‌ها می‌شود. در این مطالعه، یک چارچوب یادگیری بازنمایی نظارت‌شده به‌منظور پردازش داده‌های بالینی ناهمگون شامل ویژگی‌های عددی و دسته‌ای ارائه شده است. در این چارچوب، ویژگی‌های دسته‌ای ابتدا از طریق یک لایه تعبیه آموزش‌پذیر رمزگذاری می‌شوند، در حالی که داده‌های عددی از طریق یک لایه نرمال‌سازی پیش‌پردازش می‌شوند. سپس، داده‌های پردازش شده عددی و دسته‌ای با یکدیگر ادغام شده و توسط یک لایه اتصال کامل نگاشت می‌شوند تا بازنمایی‌های غیرخطی و موثر از مجموعه دادگان به دست آید. در نهایت، برای مقابله با مشکل عدم‌توازن کلاس‌ها و افزایش دقت در شناسایی نمونه‌های کلاس اقلیت، از یک دسته‌بند ***XGBoost*** وزن‌دهی شده استفاده شد که با تخصیص وزن‌های متفاوت به کلاس‌ها، امکان شناسایی جهش‌های نادر را فراهم می‌کند. کارایی روش پیشنهادی بر مجموعه‌داده ***NSCLC Radiogenomics*** شامل ۲۱۱ بیمار از پایگاه ***TCIA*** با اعتبارسنجی متقابل پنج فولد متوازن ارزیابی شد. روش پیشنهادی به دقت ۸۰/۹٪، حساسیت ۷۲/۲٪، اختصاصیت ۸۳/۷٪، ***F1-score*** ۶۲/۶٪، ***precision*** ۶۶/۵٪ مساحت زیر منحنی ***(AUC)*** ۰/۸۲ دست یافت. مقایسه با الگوریتم‌های موجود نشان داد که روش پیشنهادی، شناسایی جهش‌های ***EGFR*** را در داده‌های بالینی ناهمگون و نامتوازن به‌طور قابل توجهی بهبود می‌دهد.

*نویسنده‌ی مسئول

نشانی: تهران، خیابان شریعتی، پایین‌تر از پل سیدخندان، دانشگاه صنعتی خواجه نصیرالدین طوسی، دانشکده‌ی مهندسی برق

پست الکترونیک: Mansour.vali@eetd.kntu.ac.ir

۱- مقدمه

سرطان ریه شایع‌ترین سرطان تشخیص داده شده در جهان درطی چند دهه اخیر بوده است، به طوریکه در سال ۲۰۱۸، حدود ۲/۱ میلیون مورد جدید ابتلا به سرطان ریه تشخیص داده شده است که ۱۲٪ کل افراد مبتلا به سرطان را تشکیل می‌دهد [۱]. نرخ ابتلا به سرطان ریه در مردان ۱/۳۷ میلیون نفر و در زنان ۷۲۵ هزار نفر در سال ۲۰۱۸ گزارش شده است که این امر حکایت از بالا بودن نرخ ابتلا در مردان است [۲].

سرطان ریه از نظر بافت شناسی به دو دسته عمده سرطان ریه سلول‌های کوچک[1] (SCLC) و سرطان ریه سلول‌های غیر کوچک[2] (NSCLC) تقسیم می‌شود. SCLC اغلب در برونش‌ها یا راه‌های هوایی که از نای به ریه‌ها منتهی می‌شوند شروع می‌شود و سپس به ساختارهای کوچک‌تر منشعب می‌شود. پس از تأثیر بر برونش‌ها، SCLC به سرعت رشد نموده و به سایر قسمت‌های بدن از جمله مغز، کبد، و استخوان گسترش می‌یابد و منجر به مرگ تقریبا ۹۵٪ بیماران می‌شود [۳]. اصلی‌ترین علت وقوع SCLC مصرف بالای دخانیات به ویژه سیگار است. حدود ۸۰٪ الی ۸۵٪ بیماران مبتلا به سرطان ریه، از NSCLC رنج می‌برند. این نوع از سرطان ریه اغلب به کندی ایجاد می‌شود و تا زمانی که پیشرفت نکرده باشد، بدون علامت است یا علائم کمی را ایجاد می‌کند. NSCLC از لحاظ شروع درگیری سلول‌های ریه، به سه دسته عمده آدنوکارسینوم، کارسینوم سلول‌های سنگفرشی، و کارسینوم سلول‌های بزرگ تقسیم می‌شود.

آدنوکارسینوم ریه، حدود ۴۰٪ افراد مبتلا به سرطان ریه را در برگرفته و معمولا شروع آن از غدد مخاطی است [۴]. شیوع آدنوکارسینوم در بین زنان بیش از مردان بوده و افراد جوان را بیش از سایرین درگیر می‌کند. کارسینوم سلول‌های سنگفرشی عمدتاً در بیماران سیگاری رخ می‌دهد و ۲۵٪ از بیماران سرطان ریه را شامل می‌شود. حدود ۱۰٪ از بیماران مبتلا به سرطان ریه از کارسینوم سلول‌های بزرگ رنج می برند. ارتباط معناداری بین ابتلا به کارسینوم سلول‌های بزرگ با استعمال دخانیات وجود دارد [۵].

عوامل اجتماعی و اقتصادی به طور قابل توجهی بر بروز سرطان ریه تاثیر می‌گذارند، به طوری که ۵۸٪ افراد مبتلا در کشورهای با درآمد کم و متوسط شناسایی می‌شوند [۶]. ژنتیک، جنسیت، مصرف الکل و سیگار، التهاب مزمن، قرار گیری در معرض اشعه-های یونیزه کننده، قرار گرفتن در معرض آسیب‌های شغلی مرتبط با ریه، و آلودگی هوا از سایر علل ابتلا به سرطان ریه محسوب می‌شوند [۷] که از میان آن‌ها، استعمال سیگار و تنباکو از مهم‌ترین عوامل ابتلا به شمار می‌روند [۸].

روش‌های درمانی شامل جراحی، شیمی درمانی، پرتودرمانی، و درمان‌های هدفمند است. شیمی درمانی رویکرد اصلی درمان، در بیماران مبتلا به سرطان ریه است. هرچند این روش با چالش‌هایی همچون مقاومت دارویی و هدف گیری غیراختصاصی سلول‌ها مواجه است. برای غلبه بر مشکلات مذکور، از درمان‌های هدفمند استفاده می‌شود [۹]. برای درمان هدفمند، تشخیص دقیق جهش گیرنده فاکتور رشد اپیدرمی[3] (EGFR) مورد نیاز است. جهش در آکسون‌های ۱۹ و ۲۱ که در بیماران NSCLC رایج است، دامنه تیروزین کیناز[4] (TKI) را در EGFR فعال می‌کند که منجر به پاسخ خوبی برای درمان با مهارکننده‌های EGFR-TKI می‌شود. EGFR-TKI کیفیت زندگی را در بیماران بهبود می‌بخشد و عوارض جانبی شیمی درمانی استاندارد را کاهش می‌دهد.

نمونه برداری یک رویکرد معیار برای تشخیص جهش EGFR است. هرچند این روش تهاجمی بوده و با مشکلات متعددی مواجه است. ناهمگونی تومور، محلی سازی دقیق آن را پیچیده می‌کند [۱۰]. هزینه بالا، توالی یابی جهش‌های گسترده را محدود می‌کند [۱۱]. مسائلی مانند کیفیت پایین دئوکسی ریبونوکلئیک اسید[5] (DNA) و نیاز به تکرار آزمایش‌ها برای بیماران ناخوشایند است. همچنین این روش به دلیل اشتباهات نمونه گیری می‌تواند منجر به نتایج منفی کاذب شود [۱۲, ۱۳]. این محدودیت‌ها ضرورت توسعه روش‌های غیرتهاجمی، دقیق و مقرون به صرفه برای تشخیص جهش‌های EGFR در بیماران مبتلا به NSCLC را برجسته می‌کند.

دادگان بالینی یک روش رایج برای ارزیابی وضعیت بیماران NSCLC است که می‌تواند به عنوان یک ابزار قدرتمند جهت غربالگری وضعیت جهش EGFR در آن‌ها مورد استفاده قرار گیرد [۱۴-۱۷].

Yin و همکاران، از یک ماشین بردار پشتیبان[6] (SVM) با هسته خطی برای تشخیص جهش EGFR مبتنی بر داده‌های بالینی بهره گرفتند [۱۴]. در این مطالعه، اطلاعاتی شامل جنسیت، سن، محل تومور، سابقه مصرف دخانیات، اندازه تومور،

[1] Small Cell Lung Cancer
[2] Non-small Cell Lung Cancer
[3] Epidermal Growth Factor Receptor
[4] Tyrosine Kinase
[5] Deoxyribonucleic Acid
[6] Support Vector Machine

و مرحله سرطان ریه در بیماران، به عنوان متغیرهای ورودی برای آموزش و ارزیابی عملکرد الگوریتم به کار گرفته شد.

Dongو همکاران، داده‌های بالینی شامل جنسیت، سن، و وضعیت استعمال دخانیات را مورد تجزیه و تحلیل قرار دادند و از الگوریتم جنگل تصادفی[1] (RF) جهت طبقه‌بندی وضعیت جهش EGFR استفاده کردند [۱۵]. در مطالعه‌ای دیگر، Liu و همکاران، داده‌های بیماران مبتلا به آدنوکارسینوم را به منظور بررسی وضعیت جهش‌های EGFR و Kirsten rat sarcomaتحلیل نمودند [۱۶]. آن‌ها داده‌های بالینی شامل سن، جنسیت، نژاد، وضعیت استعمال دخانیات، و مرحله سرطان را جمع‌آوری کرده و با بهره‌گیری از الگوریتم RF به طبقه‌بندی این داده‌ها پرداختند. Rios و همکاران، از داده‌های کلینیکوپاتولوژیک شامل جنسیت، وضعیت استعمال دخانیات، اطلاعات آسیب شناختی، بافت‌شناسی، و مرحله تومور برای تشخیص جهش EGFR با استفاده از رگرسیون لجستیک بهره گرفتند [۱۷]. نتایج این مطالعه نشان داد که زن بودن، نداشتن سابقه استعمال دخانیات، و درجه آسیب‌شناختی پایین یا متوسط، ارتباط معناداری با بروز جهش EGFR دارند.

با وجود پیشرفت‌های قابل‌توجه در حوزه ژنومیک طی سال‌های اخیر [۱۸] و ارائه الگوریتم‌های نوین برای شناسایی جهش EGFR در بیماران مبتلا به NSCLC، توسعه الگوریتم‌هایی که علاوه بر دقت بالا، از کارایی بالینی موثر و قابل اطمینان در شرایط واقعی برخوردار باشند، همچنان یکی از چالش‌های اساسی این حوزه به‌شمار می‌رود. پردازش داده‌های بالینی به علت ناهمگونی داده‌های جمع آوری شده و نامتعادل بودن آن‌ها با چالش‌های جدی مواجه است. عمده مقالات موجود از روش‌های مرسوم رمزگذاری تک-داغ[2] یا رمزگذاری برچسب[3] استفاده می‌کنند که در استخراج الگوهای پیچیده از داده‌های بالینی ناتوان هستند. در سال‌های اخیر، یادگیری بازنمایی در داده‌های جدولی به عنوان رویکردی مؤثر برای تحلیل و مدل‌سازی داده‌ها مورد توجه فزاینده‌ای قرار گرفته است [۱۹, ۲۰]. در این حوزه، روش TabTransformer به‌عنوان یکی از رویکردهای نوین استخراج بازنمایی معرفی شده است که با بهره‌گیری از تعبیه ستونی و سازوکار ترنسفورمر، به رمزگذاری داده‌ها و استخراج همبستگی‌های میان ویژگی‌های دسته‌ای پرداخته است [۲۱]. رویکرد استخراج بازنمایی در حوزه پردازش داده‌های پزشکی نیز اهمیت ویژه‌ای دارد [۲۲]. برای نمونه، در سیستم‌های توصیه‌گر دارویی به منظور ارتقای دقت و کارایی الگوریتم‌ها، از شبکه‌ی یادگیری بازنمایی تقویت شده با دانش[4] (KERL) استفاده شده است [۲۳]. برخلاف بسیاری از مدل‌های یادگیری عمیق، KERL شبکه‌ای را ارائه می‌دهد که با بهره‌گیری از دانش حوزه‌ای بیرونی[5]، بازنمایی داده‌ها را بهبود می‌بخشد و عملکرد سیستم توصیه‌گر دارویی را به‌طور جامع ارتقا می‌دهد. علاوه بر این، در حوزه‌ی داده‌های بیان ژن نیز رویکردهای یادگیری بازنمایی مبتنی بر روش‌های خودنظارتی[6]، برای پیش‌بینی فنوتیپ توسعه یافته‌اند و نتایج امیدبخشی به همراه داشتند [۲۴].

گرچه تاکنون روش‌های متعددی برای استخراج بازنمایی از داده‌های پزشکی ارائه شده‌اند، اما توسعه‌ی رویکردی که بتواند بازنمایی‌های مؤثر و سازگار با ماهیت داده‌های جدولی را برای تشخیص جهش‌های ژنی فراهم سازد به علت ناهمگونی داده‌ها، پیچیدگی تعاملات میان ویژگی‌های بالینی عددی و دسته‌ای، و نامتعادل بودن داده‌ها، همچنان چالش برانگیز است. در این مقاله، رویکردی برای استخراج بازنمایی نظارت شده[7] (SR) ارائه شده که هدف آن رمزگذاری ویژگی‌های دسته‌ای، پردازش داده‌های عددی، و مدل‌سازی وابستگی‌های پیچیده و غیرخطی میان این دو نوع ویژگی است. در روش پیشنهادی، پس از استخراج بازنمایی از دادگان بر خلاف روش‌هایی مانند TabTransformer که از پرسپترون چند لایه برای طبقه بندی داده‌ها استفاده می‌کنند، از روش گرادیان تقویتی شدید وزن دهی شده[8] (WXGB) برای بهبود طبقه بندی داده‌های جدولی نامتوازن و استخراج دقیق‌تر و بهتر مرزهای تصمیم گیری بهره گرفته شد. در این چارچوب، برای کاهش نرخ خطای طبقه‌بندی، به نمونه‌های متعلق به کلاس اقلیت وزن بیشتری اختصاص داده می‌شود. این رویکرد موجب می‌شود که در هنگام به‌روزرسانی درخت‌های تصمیم‌گیری، مدل توجه بیشتری به نمونه‌های کلاس اقلیت معطوف کند و از غالب شدن کلاس اکثریت جلوگیری شود. به عبارت دیگر، در ساخت هر درخت تقویتی، خطای نمونه‌های کلاس اقلیت با وزن بیشتری در نظر گرفته شده و تأثیر بیشتری بر یادگیری مدل دارد، که در نتیجه مرزهای تصمیم‌گیری مناسب برای هر کلاس به تدریج شناسایی می‌شوند. همچنین، در فرآیند تجمیع[9] در WXGB، اثر نمونه‌های کلاس اکثریت تعدیل شده و مدل قادر است بدون ایجاد سوگیری، مرزهای تصمیم‌گیری صحیح را

[1] Random Forest
[2] One-hot Encoding
[3] Label Encoding
[4] Knowledge Enhanced Representation Learning
[5] External Domain Knowledge
[6] Self-supervised Learning
[7] Supervised Representation Learning
[8] Weighted eXtreme Gradient Boosting
[9] Ensemble

برای تمامی کلاس‌ها تشخیص دهد. نوآوری‌های مقاله پیشنهادی به شرح زیر می‌باشد:

- استفاده از مجموعه داده جامع برای تشخیص جهش EGFR: این مطالعه از یک مجموعه داده جامع، متشکل از ویژگی‌های جمعیت‌شناختی، نشانگرهای تصویربرداری پزشکی، و ویژگی‌های آسیب‌شناختی، در راستای تشخیص جهش EGFRبهره گرفته است.
- استخراج بازنمایی نظارت شده (SR) از مجموعه داده‌های بالینی: در این مقاله یک روش جدید جهت استخراج بازنمایی‌های موثر و کارا از مجموعه داده‌های بالینی ناهمگون ارائه شده است
- مقایسه جامع با سایر روش‌های یادگیری ماشین: به‌منظور ارزیابی عملکرد و کارایی روش پیشنهادی، مقایسه‌ای جامع با مدل‌های پایه یادگیری ماشین انجام شده است.

ادامه این مقاله به‌صورت زیر سازماندهی شده است. در بخش ۲، روش پیشنهادی، مجموعه داده‌ها، نحوه استخراج SR، طبقه بند WXGB، نحوه آموزش و ارزیابی معرفی می‌شود. بخش ۳ به گزارش نتایج، و بخش ۴ به تحلیل و بحث درخصوص آن‌ها اختصاص دارد. در نهایت، در بخش ۵ نتیجه‌گیری و افق‌های پیش‌رو بیان شده است.

۲- مواد و روش‌ها

پردازش داده‌های بالینی همواره با چالش‌های متعددی همراه بوده است. در این مطالعه، به‌طور خاص به دو چالش اساسی، یعنی ناهمگونی و عدم توازن داده‌ها، پرداخته شده است. برای مقابله با ناهمگونی داده‌ها، از استخراج SR استفاده شده است که رویکردی نوین برای بازنمایی داده‌های شامل ویژگی‌های عددی و دسته‌ای است. هدف اصلی در این بلوک، استخراج بازنمایی‌های مؤثر و مقاوم با هدف ارتقای عملکرد مدل‌های تشخیصی مبتنی بر داده‌های بالینی است. سپس، برای طبقه‌بندی بازنمایی‌های استخراج‌شده از داده‌های بالینی، از روش WXGB استفاده شده است که به طور اختصاصی برای طبقه‌بندی داده‌های نامتوازن بهینه‌سازی شده است. روندنمای کلی روش پیشنهادی در شکل ۱ نشان داده شده است.

۲-۱- پایگاه داده

در این مطالعه از مجموعه داده NSCLC Radiogenomics که در پایگاه داده TCIA در دسترس عموم قرار دارد، استفاده شده است [۲۵]. این مجموعه شامل تصاویر توموگرافی رایانه‌ای[1] (CT)، توموگرافی با گسیل پوزیترون[2] (PET)، نقشه‌های بخش‌بندی تومور، حاشیه نویسی معنایی تومورها، داده‌های بالینی، و داده‌های بیان ژن-RNA) (seq از ۲۱۱ بیمار مبتلا به NSCLC می‌باشد.

اطلاعات بالینی شامل اطلاعات جمعیت‌شناختی بیماران (سن، جنسیت، و نژاد)، سابقه استعمال دخانیات (وضعیت و تعداد بسته مصرفی در سال)، ویژگی‌های تصویری (مانند محل تومور و درصد گراند گلس)، ویژگی‌های آسیب شناختی شامل مرحله بندی تومور (تومور، ندول، دگردیسی)، اطلاعات هیستولوژی، درجه بندی تومور، و تهاجم لنفاوی است. جزئیات بیشتر از مشخصات جمعیت شناختی و بالینی بیماران در جدول ۱ خلاصه شده است.

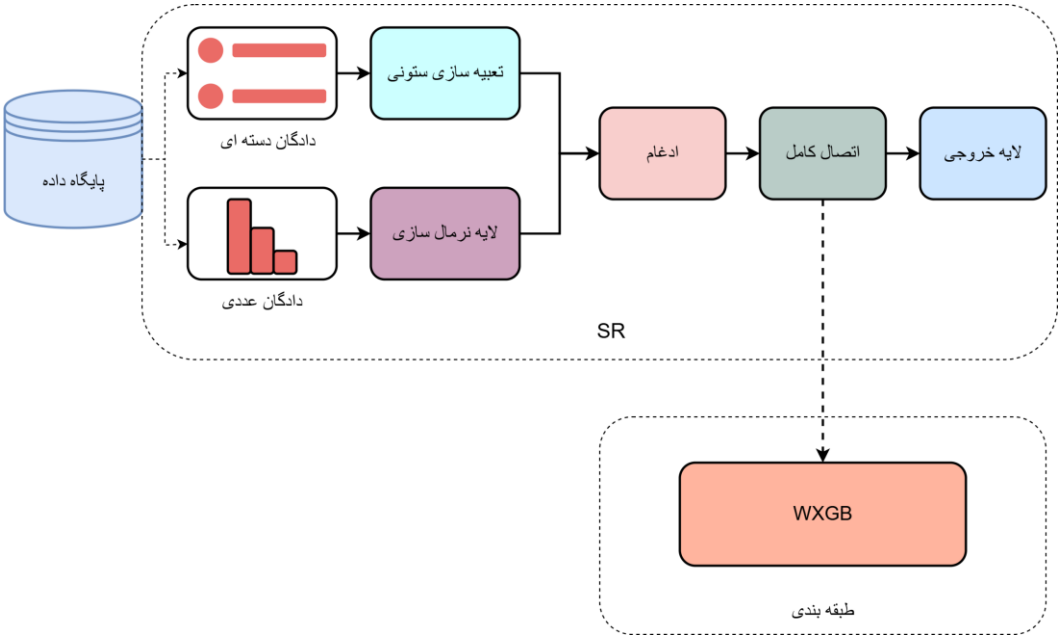
در راستای افزایش دقت تحلیل‌ها، نمونه‌هایی که وضعیت جهش ژن EGFR آن‌ها "Not collected" یا "Unknown" بود (۳۹ بیمار)، از مجموعه دادگان حذف شدند. پس از پالایش داده‌ها، مجموعه نهایی شامل اطلاعات مربوط به ۱۷۲ بیمار بود که از این میان، ۴۳ بیمار در گروه "Mutant" و ۱۲۹ بیمار در گروه"Wild-type" قرار داشتند.

۲-۲- استخراج SR

داده‌های خام معمولاً شامل نویز و اطلاعات غیرضروری هستند که می‌تواند دقت الگوریتم‌های طبقه‌بندی را کاهش دهد. یادگیری بازنمایی با توانایی استخراج ساختار و اطلاعات نهفته در داده‌ها، نقش مهمی را در بهبود دقت طبقه‌بندی ایفا می‌کند. در این مطالعه، جهت استخراج بازنمایی از داده‌های بالینی ناهمگون، یک روش یادگیری بازنمایی نظارت شده (SR) ارائه شده است. این روش برخلاف رمزگذارهای مرسوم مانند رمزگذاری تک-داغ، قادر به یادگیری روابط پیچیده و نهفته بین داده‌های دسته‌ای و عددی است. به عبارت دیگر، قادر است مشابهت‌ها و تفاوت‌های معنایی بین مقادیر دسته‌ای را در قالب بردارهای عددی بازنمایی کند، که این موضوع منجر به بهبود عملکرد مدل‌های یادگیری ماشین می‌شود. روش استخراج SR در این مطالعه شامل مراحل زیر است: تعبیه‌سازی آموزش پذیر در داده‌های دسته‌ای، استفاده از لایه نرمال‌سازی برای داده‌های عددی، ادغام داده‌های تعبیه‌شده دسته‌ای با داده‌های عددی نرمال‌شده، و در نهایت استفاده از لایه اتصال کامل برای استخراج ویژگی‌های غیرخطی و یادگیری روابط پیچیده بین ویژگی‌های عددی و دسته‌ای.

[1] Computed Tomography

[2] Positron Emission Tomography



شکل (۱)- روندنمای روش پیشنهادی جهت تشخیص جهش EGFR از داده‌های بالینی ناهمگون و نامتوازن.

فرض کنیم یک مجموعه داده $X \in R^{n\times m}$ در اختیار داریم که شامل n نمونه و m ویژگی است. در این مجموعه داده، هر نمونه $x_i \in X$ شامل ترکیبی از ویژگی‌های عددی و دسته‌ای است. بنابراین می‌توانیم هر نمونه را به صورت $x_i = (x_i^{num}, x_i^{cat})$ نمایش دهیم که در آن $x_i^{num} \in \mathbb{R}^{nc}$ شامل $nc = 3$ ویژگی عددی و $x_i^{cat} = (x_i^{(1)}, x_i^{(2)}, \dots, x_i^{(mc)})$ که در آن $mc = 17$ ویژگی دسته‌ای است.

جهت رمزگذاری ویژگی‌های دسته‌ای و پردازش آن‌ها در الگوریتم‌های یادگیری ماشین، در این مطالعه از رویکرد تعبیه آموزش پذیر [۲۱] استفاده شده است. این روش یک رویکرد آموزش پذیر است که در خلال آن یک نمایش عددی از ویژگی‌های دسته‌ای استخراج می‌شود. به طور خاص هر ویژگی دسته‌ای (x_i) با تعداد دسته‌های k_i به وسیله یک تابع تعبیه سازی $e_{\varphi_i}(x_i) \in \mathbb{R}^d$ به یک فضای برداری با ابعاد d نگاشت داده می‌شود. در نهایت بردارهای نهفته‌ی مربوط به همه‌ی ویژگی‌های دسته‌ای با $E_\varphi(x_{cat}) = \{e_{\varphi 1}(x_1), \dots, e_{\varphi m_c}(x_{m_c})\}$ نمایش داده می‌شود.

پس از تعبیه سازی داده‌های دسته‌ای، از لایه نرمال سازی[1] [۲۶]، جهت پردازش داده‌های عددی استفاده می‌شود. این روش باعث هموارتر شدن گرادیان، تسریع روند آموزش، و بهبود تعمیم پذیری مدل می‌شود. این لایه، نرمال سازی داده‌ها را روی ویژگی‌های هر نمونه به صورت جداگانه انجام می‌دهد و برای این منظور ابتدا میانگین و واریانس را بر روی داده‌ها مطابق روابط (۲،۳) محاسبه می‌کند.

$$\mu = \frac{1}{H}\sum_{j=1}^{H} x_j \quad (۲)$$

$$\sigma^2 = \frac{1}{H}\sum_{j=1}^{H}(x_j - \mu)^2 \quad (۳)$$

که در روابط فوق H بیانگر تعداد نورون‌های لایه‌ی مخفی، x_j ویژگی‌های عددی، μ مقدار میانگین، و σ^2 واریانس داده‌ها است. سپس هر ویژگی بر اساس میانگین و واریانس مطابق رابطه (۴) نرمال سازی می‌شود.

$$\hat{x_j} = \frac{x_j - \mu}{\sqrt{\sigma^2 + \varepsilon}} \quad (۴)$$

در این رابطه ε مقدار عددی بسیارکوچکی است که جهت جلوگیری از صفر شدن مخرج استفاده شده است. با استفاده از این نرمال سازی، توزیع ویژگی‌ها در هر نمونه استاندارد (میانگین صفر و واریانس واحد) می‌شود. سپس بردار ویژگی حاصل از این لایه با خروجی لایه تعبیه‌سازی ستونی ادغام می‌شود. در نهایت از یک لایه اتصال کامل جهت نگاشت غیرخطی ویژگی‌ها و یادگیری روابط پیچیده میان متغیرهای ناهمگون استفاده می‌شود. این لایه علاوه بر بهبود ظرفیت مدل در استخراج بازنمایی‌های غنی‌تر، تعمیم پذیری مدل را نیز ارتقا می‌دهد.

[1] Layer Normalization

جدول (۱)- شرح مجموعه داده های بالینی پایگاه داده NSCLC Radiogenomics.

ویژگی		تعداد بیماران
جنسیت	مرد	۱۳۵
	زن	۷۶
نژاد	آفریقایی-آمریکایی	۶
	آسیایی	۲۴
	قفقازی	۱۲۳
	اسپانیایی/لاتین	۶
	بومی هاوایی/جزایراقیانوس آرام	۳
	جمع آوری نشده	۴۹
هیستولوژی	آدنوکارسینوما	۱۷۲
	کارسینوم سلول سنگ‌فرشی	۳۵
	نامشخص	۴
مرحله تومور	T0	۶
	T1	۷۱
	T2	۵۷
	T3	۲۱
	T4	۷
	جمع آوری نشده	۴۹
مرحله ندول	N0	۱۲۹
	N1	۱۵
	N2	۱۸
	جمع آوری نشده	۴۹
مرحله	M0	۱۵۷
	M1	۵
	جمع آوری نشده	۴۹
وضعیت	غیرسیگاری	۴۸
	ترک کرده	۱۳۰
	مصرف کننده	۳۳
درجه	G1 Well differentiated	۳۲
	G2 Moderately differentiated	۷۶
	G3 Poorly differentiated	۳۳
	Other Type I	۹
	Other Type II	۱۲
	جمع آوری نشده	۴۹

۲-۳- طبقه بندی

طبقه‌بندی دقیق نمونه‌ها در مجموعه داده‌های نامتوازن، یکی از چالش‌های بنیادین در الگوریتم‌های مبتنی بر یادگیری ماشین است؛ چالشی که در دادگان با تعداد نمونه‌های کم، اهمیت و پیچیدگی بیشتری پیدا می‌کند. در داده‌های پزشکی نیز به دلیل ماهیت و شرایط خاص گردآوری، توزیع نمونه‌ها معمولاً به نفع افراد سالم بوده و سهم نمونه‌های بیمار بسیار ناچیز است [۲۷, ۲۸]. در حالی که شناسایی صحیح نمونه‌های اقلیت، نقشی حیاتی در بهبود فرآیندهای تشخیصی و درمانی ایفا می‌کند. از این رو، ارائه رویکردی که بتواند با دقت و حساسیت بالا نمونه‌های اقلیت را شناسایی کند، از اولویت‌های اساسی در تحلیل داده‌های پزشکی محسوب می‌شود.

مدل XGBoost [۲۹]، یکی از الگوریتم‌های نوین یادگیری ماشین است که با بهره گیری از روش تقویت گرادیان و بهینه سازی پیشرفته مدل، منجر به ارتقای عملکرد در مسائل طبقه بندی و رگرسیون شده است. اساس این مدل بر مبنای ساخت مدل‌های ضعیف (عمدتا درخت‌های تصمیم گیری کم عمق) بنا شده است. پس از ایجاد مدل‌های ضعیف، این الگوریتم تلاش می‌کند به صورت تدریجی، مدل را بهبود داده و خطای مدل‌های پیشین را کاهش دهد. در راستای کاهش پیچیدگی مدل و به تبع جلوگیری از بیش برازش مدل، از تنظیم کننده‌های L1 و L2 استفاده می‌شود. همچنین این مدل با بهره گیری از قابلیت پردازش موازی و بهینه سازی استفاده از حافظه، سرعت اجرای بالاتری نسبت به الگوریتم‌های مشابه دارد. قابلیت وزن‌دهی به نمونه‌های کلاس اقلیت نقشی کلیدی در کاهش سوگیری ناشی از نامتوازنی کلاس‌ها ایفا می‌کند. در این پژوهش، به‌منظور ارتقای کارایی طبقه‌بندی داده‌های نامتوازن، از نسخه‌ی وزن‌دهی‌شده‌ی الگوریتم XGBoost (WXGB) استفاده شده است. در این رویکرد، وزن کلاس اقلیت متناسب با تعداد نمونه‌های موجود در کلاس اکثریت تعیین گردید تا اثر نامتوازنی کلاس‌ها کاهش یابد.

۲-۴- فرآیند آموزش و ارزیابی مدل

در این مطالعه، به منظور بهبود تعمیم‌پذیری مدل، از روش اعتبارسنجی متقابل پنج فولد متوازن[1] استفاده شد. در این روش، داده‌های ۸۰٪ از بیماران برای آموزش مدل به کار گرفته شده و ۲۰٪ باقی‌مانده به عنوان مجموعه آزمون مورد استفاده قرار گرفت. این فرآیند به گونه‌ای تکرار شد که داده هر بیمار یک بار به عنوان مجموعه آزمون مورد استفاده قرار گرفت. در این رویکرد داده‌های هر بیمار در مجموعه آموزش یا آزمون مورد استفاده قرار گرفته و هیچگونه داده مشترکی بین دو مجموعه (آموزش و آزمون) وجود ندارد. بدین ترتیب، استحکام مدل، به ویژه در مواجهه با نمونه‌های جدید، افزایش می‌یابد.

ابتدا بخش SR در هر فولد با استفاده از داده‌های آموزشی و در چارچوب یک فرایند یادگیری نظارت شده آموزش داده شد.

[1] Stratified K-fold

سپس در مرحله‌ی آزمون، مدل آموزش‌دیده با داده‌های آموزش برای استخراج بازنمایی از داده‌های آزمون به‌کار گرفته شد. به بیان دیگر، در فرایند آموزش، داده‌های ورودی به همراه برچسب‌های متناظر جهت یادگیری بازنمایی به مدل ارائه می‌شوند. پس از اتمام مرحله‌ی آموزش، بازنمایی استخراج‌شده از لایه‌ی پیش از خروجی به عنوان ویژگی‌های نهایی مورد استفاده قرار می‌گیرد. به عبارتی داده‌های ابعاد ورودی شبکه ۲۰×۱ است که شامل ۲۰ ویژگی عددی و دسته‌ای می‌باشد و بازنمایی استخراج‌شده ابعادی برابر با ۳۲×۱ دارد؛ به عبارتی ویژگی‌ها به فضایی با ابعاد بالاتر نگاشت می‌شوند. برای آموزش از تابع هزینه Cross-Entropy Sparse Categorical و بهینه‌ساز Adam با مومنتوم Nesterov [۳۰] استفاده شده است. سپس برای ارزیابی مدل، داده‌های آزمون (فقط ویژگی‌های بیماران بدون برچسب) که قبلاً توسط مدل مشاهده نشده‌اند، به‌عنوان ورودی به شبکه اعمال شده و بازنمایی مربوطه استخراج می‌شود. پس از استخراج بردارهای بازنمایی از داده‌های مورد پردازش از الگوریتم WXGB که برای طبقه بندی داده‌های نامتوازن بهینه سازی شده است استفاده می‌شود. به این ترتیب ورودی مدل WXGB یک بردار بازنمایی با ابعاد ۳۲×۱ بوده و خروجی آن به صورت دو کلاسه است. در این مطالعه وزن‌دهی کلاس اقلیت در الگوریتم WXGB، بر طبق نسبت تعداد نمونه‌های کلاس اکثریت به نمونه‌های کلاس اقلیت انجام می‌پذیرد. همچنین لازم به ذکر است سایر فراپارامترهای مدل پیشنهادی مطابق جدول ۲ بهینه سازی شده‌اند.

جدول (۲)- فراپارامترهای بهینه مدل پیشنهادی.

بلوک	فراپارامتر	مقدار بهینه
تعبیه ستونی	ابعاد نگاشت	۴
اتصال کامل	تعداد لایه‌ها	۱
	تعداد نورون‌ها	۳۲
	تابع فعالسازی	ELU
طبقه بند	نرخ یادگیری	۰/۰۳
	عمق بیشینه	۳
	تعداد تخمین‌ها	۵۰
	Min child weight	۴
	گاما	۰/۰۱
	زیرنمونه‌گیری	۰/۸

۳- نتایج

برای ارزیابی عملکرد روش پیشنهادی، از معیارهای دقت، حساسیت، اختصاصیت، F1-score، Precision، و سطح زیر منحنی[1] (AUC) استفاده شد که برای ارزیابی عملکرد روش‌های طبقه‌بندی به ویژه در مجموعه دادگان نامتعادل مهم هستند. این معیارها با رویکرد روش پیشنهادی (SR-EGFR) ، به ترتیب ۸۰/۹۱٪، ۷۲/۲۲٪، ۸۳/۷۵٪، ۶۲/۶۲٪، ۶۶/۵۱٪ و ۰/۸۲ به دست آمد.

به‌منظور ارزیابی نقش بلوک SR در پردازش داده‌های بالینی، نتایج روش پیشنهادی با مدل پایه (قبل از استخراج SR) مقایسه گردید. عملکرد مدل پایه بر اساس داده‌های خام به ترتیب شامل دقت ۷۰/۹۲٪، حساسیت ۵۵/۸۳٪، اختصاصیت ۷۶/۰۶٪، F1-score ۴۴/۴۷٪، Precision ۴۸/۷۲٪ و AUC ۰/۷۴ بود. استفاده از بلوک SR موجب بهبود قابل توجهی در نتایج گردید؛ به گونه‌ای که دقت ۱۰٪، حساسیت ۱۶٪، اختصاصیت ۸٪، F1-score ۱۸٪، Precision ۱۸٪ و AUC ۰/۰۸ افزایش پیدا کرد.

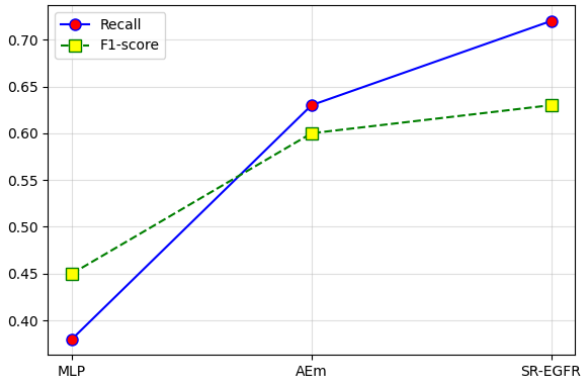
جهت ارزیابی معناداری بهبود عملکرد، آزمون آماری t-test بر روی مقادیر AUC قبل و پس از استخراج SR انجام شد. نتایج نشان داد که استفاده از SR موجب بهبود معنادار عملکرد مدل ($p\text{-value} < 0.05$) گردیده است.

برای ارزیابی دقیق‌تر ساختار پیشنهادی، دو مدل مقایسه‌ای دیگر نیز توسعه داده شد. در مدل نخست، از ابعاد تعبیه‌سازی تطبیقی[2] (AEm) که بُعد تعبیه‌سازی هر ویژگی دسته‌ای را متناسب با تعداد دسته‌های آن تعیین می‌کند، استفاده شده است. مدل دوم با هدف بررسی نقش استخراج SR و تأثیر استفاده از روش یادگیری تجمیعی در طبقه‌بندی داده‌های EGFR توسعه یافته است. در این مدل، از همان شبکه عصبی مورد استفاده در ساختار پیشنهادی برای استخراج ویژگی‌های SR بهره گرفته شد، با این تفاوت که شبکه به‌صورت انتها به انتها[3] آموزش داده شد تا هم استخراج ویژگی و هم طبقه‌بندی را به‌طور مستقیم انجام دهد. نتایج نشان‌دهنده کارایی بالای روش SR-EGFR در طبقه‌بندی صحیح نمونه‌های کلاس ’Mutant‘ است. مقایسه‌ای میان حساسیت و F1-score روش پیشنهادی و سایر مدل‌های توسعه داده‌شده در این بخش، در شکل ۲ نمایش داده شده است.

الگوریتم XGBoost قادر به استخراج الگوهای پیچیده از داده‌ها حتی در مجموعه دادگان با تعداد نمونه‌های کم است،

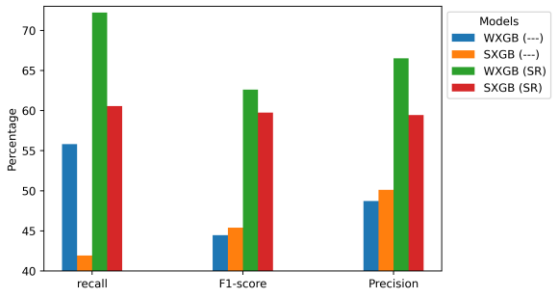
[1] Area Under the Curve
[2] Adaptive Embedding
[3] End-to-end

در حالیکه شبکه عصبی برای یادگیری همین الگوها نیاز به حجم زیادی از داده‌ها دارد. علاوه بر این، در مدل AEm، به‌منظور یادگیری مؤثر، ابعاد نگاشت بزرگ‌تری مورد نیاز است که توان استخراج ویژگی‌های موثر را از دادگان دسته‌ای داشته باشد.



شکل (۲) – مقایسه حساسیت و F1-score مدل پیشنهادی با AEm و شبکه عصبی.

جهت بررسی دقیق‌تر تاثیر وزن دهی XGBoost (WXGB) و اهمیت آن، به جای وزندهی روش XGBoost از روش نمونه‌سازی افزایشی اقلیت با داده‌های مصنوعی[1] (SMOTE) و ترکیب آن با الگوریتم XGBoost (SXGB) پیش و پس از استخراج SR استفاده نموده و نتایج را در شکل ۳ مورد مقایسه قرار دادیم.



شکل (۳)- مقایسه حساسیت، F1-score، و Precision روش WXGB با SXGB قبل و پس از استخراج بازنمایی.

همان‌طور که در شکل ۳ نشان داده شده است، وزن‌دهی در الگوریتم XGBoost (WXGB) توانایی شناسایی کلاس‌های اقلیت را نسبت به SMOTE (SXGB) بهبود می‌بخشد. در مجموعه داده‌هایی با تعداد محدود نمونه، الگوریتم SMOTE ممکن است دچار بیش‌برازش شود، زیرا نمونه‌های مصنوعی تولید شده نمی‌توانند تنوع واقعی داده‌ها را به طور کامل بازتاب دهند. در مقابل، وزن‌دهی نمونه‌ها در XGBoost باعث می‌شود که نمونه‌های کلاس اقلیت تأثیری معادل نمونه‌های کلاس اکثریت در فرآیند یادگیری داشته باشند، که این امر منجر به بهبود عملکرد در مقایسه با SMOTE می‌گردد.

۴- بحث و بررسی

تشخیص دقیق جهش EGFR جهت ارائه درمان‌های هدفمند در بیماران مبتلا به NSCLC ضروری است. الگوریتم‌های یادگیری ماشین دریچه ارزشمندی جهت شناسایی خودکار جهش EGFR گشوده‌اند. هر چند این الگوریتم‌ها عمدتا به دادگان حجیم جهت آموزش و تعمیم پذیری مدل‌ها نیاز دارند که با توجه به ماهیت جهش EGFR در بیماران NSCLC، جمع آوری آن هزینه‌بر و زمان‌بر است. لذا ارائه روشی که بتواند الگوهای جهش EGFR را از دادگان کم، نامتوازن، و ناهمگون شناسایی کند، اهمیت دوچندانی دارد.

۴-۱- مقایسه با الگوریتم های یادگیری ماشین

به‌منظور ارزیابی دقیق‌تر کارایی چارچوب پیشنهادی، مجموعه‌ای از الگوریتم‌های یادگیری ماشین شامل k-نزدیکترین همسایگی[2] (KNN)، درخت تصمیم گیری وزن‌دهی شده[3] (WDT)، جنگل تصادفی وزن‌دهی شده[4] (WRF)، تقویت گرادیان[5] (GB)، CatBoost وزن‌دهی شده[6] (WCB)، و WXGB به‌عنوان مدل‌های پایه انتخاب شدند. این الگوریتم‌ها یک‌بار بر روی داده‌های خام و بار دیگر پس از استخراج SR مورد بررسی قرار گرفتند. نتایج حاصل در جدول ۳ ارائه شده است.

در میان الگوریتم‌های مختلف یادگیری ماشین، الگوریتم KNN کمترین میزان حساسیت را نشان داد. این الگوریتم که بر مبنای نزدیک‌ترین همسایگی نمونه‌ها عمل می‌کند، به دلیل رویکرد خود و توزیع نامتوازن داده‌ها—که عمدتاً شامل نمونه‌های سالم است—خطای طبقه‌بندی بالایی دارد. با این حال، پس از استخراج بازنمایی از داده‌ها، حساسیت مدل به‌طور قابل توجهی (حدود ۳۰٪) بهبود یافت، که نشان‌دهنده اثربخشی روش پیشنهادی در مواجهه با داده‌های بالینی ناهمگون و نامتوازن است.

نتایج حاکی از آن است که روش WDT در مقایسه با که قابلیت تفکیک دقیق نمونه‌های اقلیت، حتی با تعداد نمونه‌های

[1] Synthetic Minority Oversampling Technique
[2] K-nearest Neighbor
[3] Weighted Decision Tree
[4] Random Forest
[5] Gradient Boosting
[6] Weighted CatBoost

کم را فراهم می‌کند. این رویکرد باعث بهبود حساسیت مدل می‌شود. هرچند احتمال بیش برازش مدل را افزایش می‌دهد که می‌تواند منجر به کاهش اختصاصیت و دقت کلی مدل شود. بررسی سایر معیارهای ارزیابی نشان می‌دهد که هرچند روش WDTاز نظر حساسیت عملکرد مناسبی دارد، اما ضعف آن در سایر معیارها منجربه محدودیت‌هایی در کاربردهای واقعی می‌شود. پس از استخراج SR از داده‌ها و اعمال آن به مدل WDT، بهبود قابل توجهی در معیارهای حساسیت، دقت، اختصاصیت،F1-Score و AUC به ترتیب در حدود ۸٪، ۵٪، ۹٪، ۸٪، و ۰/۰۴ نسبت به استفاده از داده‌های خام مشاهده شد. سپس انواع روش‌های مبتنی بر درخت تصمیم گیری شامل RF، GB، CatBoost، وXGBoost توسعه داده شده و مورد مقایسه قرار گرفت. بالاترین میزان اختصاصیت (۸۹/۱۷٪) توسط الگوریتم GB به دست آمد، هرچند حساسیت این الگوریتم (۴۰/۲۸٪) پایین بوده و قدرت بالایی در شناسایی نمونه‌های اقلیت ندارد. استفاده از SR، گرچه حدود ۶٪ حساسیت این مدل را بهبود می‌دهد، اما همچنان حساسیت پایین، کارکرد آن را روی داده‌های واقعی محدود می‌کند.

جدول (۳)- مقایسه نتایج الگوریتم پیشنهادی با سایر الگوریتم های یادگیری ماشین پیش و پس از استخراج SR.

مدل	نوع ویژگی	دقت	حساسیت	اختصاصیت	Precision	F1-score	AUC
KNN	---	۷۱/۵۳٪±۵/۳۸٪	۲۳/۶۱٪±۱۵/۶۵٪	۸۷/۶۳٪±۸/۳۰٪	۴۰/۵۶٪±۱۶/۱۵٪	۲۷/۷۵٪±۱۲/۶۱٪	۰/۵۹±۰/۰۴
	SR	۷۸/۰۰٪±۷/۷۵٪	۵۳/۶۱٪±۶/۷۸٪	۸۶/۰۶٪±۸/۸۳٪	۵۹/۸۸٪±۱۶/۹۴٪	۵۵/۸۳٪±۹/۷۶٪	۰/۷۷±۰/۰۵
WDT	---	۶۷/۵۳٪±۸/۳۷٪	۵۸/۳۳٪±۸/۷۹٪	۷۰/۷۱٪±۱۳/۳۷٪	۴۳/۱۱٪±۱۳/۷۰٪	۴۷/۷۹٪±۴/۹۱٪	۰/۶۹±۰/۰۸
	SR	۷۶/۷۹٪±۵/۹۲٪	۶۷/۲۲٪±۱۰/۸۳٪	۷۹/۸۵٪±۷/۳۹٪	۵۳/۵۸٪±۸/۶۷٪	۵۹/۲۷٪±۸/۴۳٪	۰/۷۵±۰/۰۷
WRF	---	۷۱/۵۱٪±۶/۶۱٪	۴۶/۹۴٪±۱۹/۸۰٪	۷۹/۹۱٪±۶/۷۴٪	۴۳/۲۹٪±۱۱/۳۲٪	۴۴/۱۲٪±۱۴/۷۳٪	۰/۷۶±۰/۰۹
	SR	۷۹/۷۳٪±۷/۰۷٪	۶۰/۵۶٪±۴/۸۷٪	۸۶/۰۹٪±۷/۹۲٪	۶۱/۲۳٪±۱۳/۳۲٪	۶۰/۵۹٪±۹/۱۴٪	۰/۸۱±۰/۰۹
GB	---	۷۶/۷۹٪±۶/۶۵٪	۴۰/۲۸٪±۲۱/۷۰٪	**۸۹/۱۷٪±۳/۱۵**٪	۵۳/۰۰٪±۱۳/۰۴٪	۴۴/۸۵٪±۱۸/۵۸٪	۰/۷۵±۰/۰۵
	SR	۷۶/۸۱٪±۵/۴۷٪	۴۶/۳۹٪±۱۰/۳۷٪	۸۶/۸۶٪±۶/۳۷٪	۵۵/۲۸٪±۱۲/۹۳٪	۵۰/۰۰٪±۱۰/۱۸٪	۰/۷۹±۰/۰۵
WCB	---	۷۲/۱۳٪±۴/۰۱٪	۵۸/۶۱٪±۱۱/۲۱٪	۷۶/۸۰٪±۵/۲۲٪	۴۵/۸۷٪±۴/۷۳٪	۵۱/۰۳٪±۵/۷۲٪	۰/۷۵±۰/۰۵
	SR	۷۹/۱۵٪±۷/۱۵٪	۶۵/۰۰٪±۲/۲۸٪	۸۳/۷۸٪±۹/۸۹٪	۶۰/۲۵٪±۱۳/۶۵٪	۶۱/۷۷٪±۷/۱۹٪	۰/۸۱±۰/۰۸
WXGB	---	۷۰/۹۲٪±۸/۲۷٪	۵۵/۸۳٪±۱۹/۰۰٪	۷۶/۰۶٪±۸/۱۷٪	۴۸/۷۲٪±۱۳/۹۵٪	۴۴/۴۷٪±۱۳/۴۵٪	۰/۷۴±۰/۰۹
	SR	**۸۰/۹۱٪±۸/۵۶**٪	**۷۲/۲۲٪±۵/۱۹**٪	۸۳/۷۵٪±۹/۹۰٪	**۶۶/۵۱٪±۱۰/۹۷**٪	**۶۲/۶۲٪±۱۶/۲۱**٪	**۰/۸۲±۰/۰۶**

بالاترین میزان حساسیت در داده‌های خام (۵۸/۶۱٪) و F1-score (۵۱/۰۳٪) توسط الگوریتم WCB به دست آمد. شایان ذکر است که ترکیب متقارن درخت‌ها در این الگوریتم، نقش مهمی در بهبود شناسایی نمونه‌های اقلیت در مسائل با داده‌های نامتوازن ایفا می‌کند. روش XGBoost بالاترین عملکرد را در میان معیارهای اصلی طبقه‌بندی، شامل دقت، حساسیت، F1-score، Precision، و AUC، هنگام طبقه‌بندی ویژگی‌های استخراج شده از SR داشت.

نتایج نشان داد که استفاده از SR موجب بهبود قابل توجه عملکرد تمامی الگوریتم‌ها در معیارهای مختلف ارزیابی گردیده است. بیشترین افزایش دقت و اختصاصیت مربوط به WXGB بود که به ترتیب حدود ۱۰٪ و ۸٪ افزایش یافت، در حالیکه بیشترین افزایش حساسیت ،F1-score، Precision، و AUC مربوط به KNN بود که به ترتیب ۳۰٪، ۲۸٪، ۱۹٪ و ۰/۱۸ افزایش یافت. برای مقایسه بهتر، در شکل ۴ مقایسه‌ای بین حساسیت و AUC الگوریتم‌های مختلف یادگیری ماشین پیش و پس از استخراج SR ارائه شده است.

۲-۴- مطالعه حذف اجزا

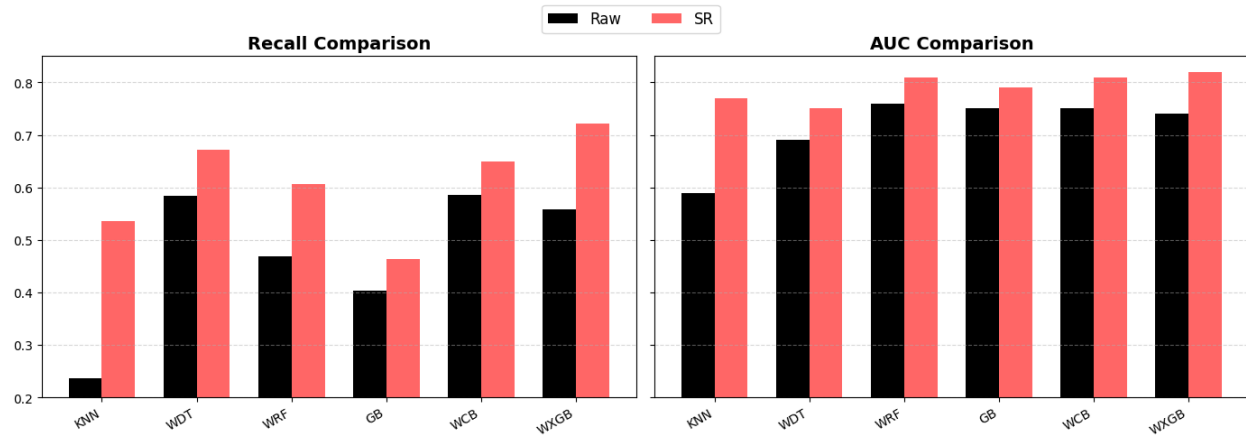
سپس جهت بررسی جامع‌تر و دقیق‌تر نقش مؤلفه‌های مختلف روش پیشنهادی در بهبود عملکرد کلی، مطالعه‌ای تحلیلی به‌صورت حذف اجزا[1] انجام شد. در این مطالعه، روش پیشنهادی به‌طور جداگانه با استفاده از بلوک تعبیه‌سازی ستونی، لایه نرمال‌سازی، و لایه اتصال کامل ارزیابی گردید. نتایج به‌دست‌آمده در جدول ۴ ارائه شده است.

مشاهده شد که حساسیت و AUC با افزودن لایه اتصال کامل به ترتیب حدود ۱۰٪ و ۰/۲۲ نسبت به مدل پایه بهبود یافته‌اند. بهبود در AUC همچنین با افزودن لایه نرمال‌سازی جهت

[1] Ablation Study

پردازش داده‌های عددی نیز مشاهده شد. با افزودن لایه تعبیه‌سازی ستونی، دقت حدود ۹٪، اختصاصیت ۱۰٪، F1-score ۳۱٪، Precision ۱۱٪، و AUC به میزان ۰/۱۹ نسبت به مدل پایه بهبود یافته است. این نتایج نشان‌دهنده ارتقای عملکرد قابل توجه این روش در تحلیل داده‌های بالینی است؛ با توجه به اینکه عمده داده‌ها در این مطالعه از نوع دسته‌ای هستند، این امر را می‌توان به کارایی روش رمزنگاری در بررسی و تحلیل داده‌های دسته‌ای نسبت داد.

در ادامه، ترکیب‌های مختلف شامل هر دو لایه ممکن از میان سه لایه موجود به‌طور جداگانه مورد ارزیابی قرار گرفت؛ به عبارت دیگر، در هر ترکیب، دو لایه حفظ شده و لایه سوم حذف گردید تا تأثیر هر یک بر عملکرد کلی بررسی شود. افزودن همزمان دو لایه نرمال‌سازی داده‌ها و اتصال کامل، نسبت به استفاده منفرد از هر یک از این لایه‌ها، نتوانست بهبودی را در کارایی مدل ایجاد کند. افزایش تعداد لایه‌ها بدون افزودن اطلاعات معنایی جدید، پیچیدگی مدل و خطر بیش برازش را افزایش می‌دهد. این در حالی است که افزودن لایه تعبیه‌سازی ستونی در ترکیب با هر یک از لایه‌های دیگر، مانند لایه نرمال‌سازی و اتصال کامل، منجر به بهبود عملکرد مدل نسبت به استفاده منفرد از هر یک از این لایه‌ها می‌شود.



شکل (۴)- مقایسه الف) AUC و ب) حساسیت الگوریتم پیشنهادی با سایر الگوریتم های یادگیری ماشین پیش و پس از استخراج SR.

در مقایسه با استفاده از دو لایه اتصال کامل و تعبیه سازی ستونی استفاده از لایه تعبیه سازی ستونی و لایه اتصال کامل به بهبود همه معیارهای ارزیابی منجر می‌شود. استفاده از لایه نرمال‌سازی گرچه نوسانات یادگیری را کاهش می‌دهد اما همبستگی های غیرخطی میان داده‌های عددی و دسته‌ای را نمی‌تواند مدل کند. لذا با افزودن لایه اتصال کامل و نگاشت دادگان به فضای غیرخطی اطلاعات معنایی مفیدی استخراج می‌شود که تا حدی عملکرد مدل را نسبت به استفاده از لایه‌های نرمال سازی و تعبیه سازی ستونی بهبود می‌دهد. در نهایت، افزودن همه لایه‌ها با در نظر گرفتن تعبیه ستونی، نرمال‌سازی و لایه اتصال کامل موجب می‌شود که علاوه بر استخراج اطلاعات مفید برای طبقه‌بندی داده‌ها، نوسانات داخلی شبکه در بخش داده‌های عددی توسط لایه نرمال‌سازی کنترل شده و نتایج بهبود قابل توجهی نسبت به سایر حالات داشته باشد.

۳-۴- مقایسه با روش های موجود

در مطالعه Yin و همکاران [۱۴]، پایگاه داده‌ای از ۳۰۱ بیمار جمع آوری شده و سپس بیماران موجود در این پایگاه داده به دو بخش آموزش و ارزیابی به ترتیب با ۱۹۸ و ۱۰۳ بیمار، تقسیم شد. دقت و AUC این مطالعه به ترتیب به ۵۹/۲۲٪ و ۰/۶۴ رسید. این مدل فقط یکبار روی مجموعه محدودی از داده‌ها آموزش دیده و روی بخشی از دادگان مورد ارزیابی قرار گرفته است که تعمیم پذیری مدل را برای نمونه‌های جدید زیر سوال می‌برد. همچنین در این مدل از تعداد محدودی از ویژگی‌ها (شامل جنسیت، سن، محل تومور، سابقه مصرف دخانیات، اندازه تومور، و مرحله سرطان ریه در بیماران) مورد استفاده قرار گرفته است و از مجموعه‌ی کامل‌تری از مشخصه‌های بالینی و پارامترهای مولکولی یا پاتولوژیک استفاده نشده است که منجر به کاهش عملکرد مدل در شناسایی نمونه‌های 'Wild-type' از نمونه‌های 'Mutant' شده است.

Dong و همکاران [۱۵]، ۵۲۵ بیمار را مورد ارزیابی قرار دادند که در طی این مطالعه ۱۶۳ بیمار برای ارزیابی استفاده شد. دقت، حساسیت، اختصاصیت، و AUC، به ترتیب به ۶۹/۲۵٪، ۶۶/۷۲٪، ۷۸/۶۴٪، و ۰/۷۱ رسید. این مطالعه نیز بخشی از دادگان را برای آموزش استفاده نموده و مجموعه داده‌های متعادلی را در فرآیند آموزش ایجاد کرده است در حالیکه در مجموعه داده‌های آزمون بخش محدودی از داده‌ها (حدود ۱۵٪) در دسته 'Mutant' قرار داشته و عمده بیماران در گروه 'Wild-type' قرار داشتند که می‌تواند برخی معیارهای ارزیابی را به صورت کاذب افزایش دهد و تعمیم پذیری مدل را تحت الشعاع قرار دهد. در این مطالعه تنها تعداد محدودی از ویژگی‌های بالینی، شامل جنسیت، سن و وضعیت استعمال دخانیات، مورد استفاده قرار گرفته‌اند که ممکن است اطلاعات کافی درباره وضعیت جهش ژنی بیماران فراهم نکنند. در مطالعه Liu و همکاران [۱۶]، ۲۹۸ بیمار دارای آدنوکارسینوم ریه مورد بررسی قرار گرفته و AUC روش پیشنهادی به ۰/۶۷ رسید. در این مطالعه با توجه به تعداد محدود بیماران، از روش k-fold هم استفاده نشده است که تعمیم پذیری آن را تحت الشعاع قرار می‌دهد. در مطالعه Rios و همکاران [۱۷]، ۷۶۳ بیمار دارای آدنوکارسینوم ریه مورد بررسی قرار گرفته که از این میان ۳۵۳ بیمار برای آموزش مدل و سایر بیماران جهت ارزیابی مورد استفاده قرار گرفتند، AUC این روش به ۰/۶۷ رسید. در این مطالعه، تنها پنج ویژگی بالینی مورد استفاده قرار گرفته و مدل بر روی بخشی از داده‌ها آموزش داده شده و بر روی بخش دیگری ارزیابی شد. محدودیت در تعداد ویژگی‌های بالینی و حجم محدود داده‌ها، موجب نگرانی‌هایی در خصوص تعمیم‌پذیری مدل می‌شود.

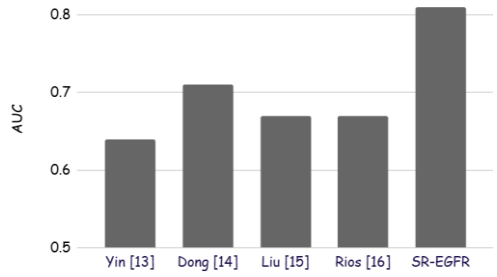
جدول (۴)- بررسی تاثیر لایه های مختلف SR بر روی عملکرد تشخیص جهش EGFR در بیماران مبتلا به سرطان ریه سلول‌های غیر کوچک.

AUC	F1-score	Precision	اختصاصیت	حساسیت	دقت	اتصال کامل	نرمال سازی	تعبیه ستونی
۰/۵۹±۰/۰۴	٪۲۷/۷۵±٪۱۲/۶۱	٪۴۸/۷۲±٪۱۳/۹۵	٪۷۶/۰۶±٪۸/۱۷	٪۵۵/۸۳±٪۱۹/۰۰	٪۷۰/۹۲±٪۸/۲۷	×	×	×
۰/۸۱±۰/۰۶	٪۶۰/۹۵±٪۱۲/۸۴	٪۵۷/۵۰±٪۱۳/۶۶	٪۸۲/۹۸±٪۷/۹۰	٪۶۵/۵۶±٪۱۳/۱۵	٪۷۸/۵۷±٪۸/۳۵	✓	×	×
۰/۷۴±۰/۰۹	٪۴۸/۷۳±٪۱۳/۹۵	٪۴۴/۴۷±٪۱۳/۴۵	٪۷۶/۰۶±٪۸/۱۷	٪۵۵/۸۳±٪۱۹/۰۰	٪۷۰/۹۲±٪۸/۲۷	×	✓	×
۰/۷۸±۰/۰۶	٪۵۸/۳۰±٪۵/۰۹	٪۵۹/۳۳±٪۸/۱۰	**٪۸۶/۰۴±٪۵/۸۸**	٪۵۸/۰۶±٪۶/۵۴	٪۷۹/۰۹±٪۴/۱۶	×	×	✓
۰/۷۴±۰/۰۹	٪۴۸/۷۳±٪۱۳/۹۵	٪۴۴/۴۷±٪۱۳/۴۵	٪۷۶/۰۶±٪۸/۱۷	٪۵۵/۸۳±٪۱۹/۰۰	٪۷۰/۹۲±٪۸/۲۷	✓	✓	×
۰/۷۹±۰/۰۸	٪۶۱/۷۲±٪۱۱/۴۳	٪۵۷/۵۹±٪۱۱/۸۱	٪۸۲/۹۵±٪۶/۴۰	٪۶۷/۲۲±٪۱۳/۳۸	٪۷۹/۱۳±٪۶/۴۰	×	✓	✓
۰/۸۰±۰/۱۰	٪۶۴/۸۷±٪۱۱/۰۵	٪۶۱/۵۲±٪۱۵/۱۶	٪۸۳/۷۸±٪۹/۸۹	٪۶۹/۷۲±٪۶/۴۰	٪۸۰/۳۳±٪۸/۴۸	✓	×	✓
۰/۸۲±۰/۰۶	**٪۶۲/۶۲±٪۱۶/۲۱**	**٪۶۶/۵۱±٪۱۰/۹۷**	٪۸۳/۷۵±٪۹/۹۰	**٪۷۲/۲۲±٪۵/۱۹**	**٪۸۰/۹۱±٪۸/۵۶**	✓	✓	✓

به طور کلی روش‌های پیشین از نظر عملکرد مدل و کارایی در سیستم‌های تشخیص پزشکی دچار محدودیت‌های جدی هستند. بسیاری از این روش‌ها، تنها یکبار مدل را آموزش داده و بخشی از دادگان را جهت ارزیابی عملکرد استفاده نموده اند، که تعمیم پذیری مدل را برای نمونه‌های جدید زیر سوال می‌برد. همچنین در روش‌های پیشین عمدتا از ویژگی‌های محدودی جهت شناسایی و تشخیص جهش EGFR استفاده شده که منجر به عملکرد ضعیف و در برخی مواقع سوگیری مدل‌ها می‌شود.

برخلاف روش‌های پیشین، ما در این مقاله از رویکرد اعتبار سنجی متقاطع متوازن با ۵ فولد استفاده نمودیم که تضمینی جهت تعمیم پذیری مدل ارائه می‌دهد. همچنین از یک پایگاه داده با ۲۰ ویژگی از بیماران استفاده نمودیم که اطلاعات کامل‌تری را در مورد وجوه مختلف EGFR را فراهم می‌کند.

عملکرد روش پیشنهادی با الگوریتم‌های پیشین مورد مقایسه قرار گرفته و در شکل ۵ نمایش داده شده است. مقایسه نتایج الگوریتم SR-EGFR با سایر الگوریتم‌های مطرح در این زمینه، بیانگر بهبود قابل توجه AUC نسبت به سایر روش‌ها است.



شکل (۵)- مقایسه AUC روش پیشنهادی با سایر الگوریتم های مطرح در تشخیص جهش EGFR.

۴-۴- محدودیت ها و پیشنهادات

با وجود پیشرفت‌های نشان داده شده در این مطالعه، همچنان محدودیت‌هایی جهت تشخیص جهش EGFR با یک مدالیته وجود دارد. ترکیب انواع دیگر دادگان مانند دادگان CT، پروفایل‌های ژنومی، یا نشانگرهای زیستی مولکولی، می‌تواند درک جامع‌تری از الگوهای جهش و بهبود دقت پیش بینی را فراهم کند. لذا تحقیقات آینده باید دادگانی را با حجم بیشتر گردآوری نموده و با ادغام دیگر مدالیته‌های جمع آوری داده، برای توسعه مدل‌های جامع تشخیص جهش EGFR در بیماران NSCLC گام بردارد.

۵- نتیجه گیری

طراحی و توسعه سیستم‌های تشخیصی و درمانی مبتنی بر هوش مصنوعی برای ارزیابی وضعیت بیماران مبتلا به سرطان ریه از اهمیت ویژه‌ای برخوردار است. در این مطالعه، روش نوین SR-EGFR برای تشخیص خودکار و دقیق جهش EGFR در بیماران مبتلا به NSCLC با استفاده از پردازش داده‌های بالینی توسعه یافت. این روش با ایجاد بازنمایی نظارت‌شده از داده‌های ناهمگون و بهره‌گیری از مدل بهینه‌شده WXGB، عملکرد قابل‌توجهی در طبقه‌بندی داده‌های نامتوازن و ناهمگون نشان داد. نتایج حاصل، راهکاری نوین و مؤثر برای تشخیص جهش EGFR ارائه داده است که افق‌های جدیدی را در انتخاب شیوه درمان و بهبود کیفیت زندگی بیماران، گشوده است.

مراجع

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: a cancer journal for clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[2] M. B. Schabath and M. L. Cote, "Cancer progress and priorities: lung cancer," *Cancer epidemiology, biomarkers & prevention*, vol. 28, no. 10, pp. 1563–1579, 2019.

[3] R. Meuwissen, S. C. Linn, R. I. Linnoila, J. Zevenhoven, W. J. Mooi, and A. Berns, "Induction of small cell lung cancer by somatic inactivation of both Trp53 and Rb1 in a conditional mouse model," *Cancer cell*, vol. 4, no. 3, pp. 181–189, 2003.

[4] E. Kuhn, P. Morbini, A. Cancellieri, S. Damiani, A. Cavazza, and C. E. Comin, "Adenocarcinoma classification: patterns and prognosis," *Pathologica*, vol. 110, no. 1, pp. 5–11, 2018.

[5] J. E. Muscat, S. D. Stellman, Z.-F. Zhang, A. I. Neugut, and E. L. Wynder, "Cigarette smoking and large cell carcinoma of the lung," *Cancer epidemiology, biomarkers & prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, vol. 6, no. 7, pp. 477–480, 1997.

[6] D. M. Parkin, F. Bray, J. Ferlay, and A. Jemal, "Cancer in africa 2012," *Cancer Epidemiology, Biomarkers & Prevention*, vol. 23, no. 6, pp. 953–966, 2014.

[7] J. Malhotra, M. Malvezzi, E. Negri, C. La Vecchia, and P. Boffetta, "Risk factors for lung cancer worldwide," *European Respiratory Journal*, vol. 48, no. 3, pp. 889–902, 2016.

[8] A. Leiter, R. R. Veluswamy, and J. P. Wisnivesky, "The global burden of lung cancer: current status and future trends," *Nature reviews Clinical oncology*, vol. 20, no. 9, pp. 624–639, 2023.

[9] G. Liang, W. Fan, H. Luo, and X. Zhu, "The emerging roles of artificial intelligence in cancer drug development and precision therapy," *Biomedicine & Pharmacotherapy*, vol. 128, p. 110255, 2020.

[10] J.-Y. Han *et al.*, "First-SIGNAL: first-line single-agent iressa versus gemcitabine and cisplatin trial in never-smokers with adenocarcinoma of the lung," *Journal of clinical oncology*, vol. 30, no. 10, pp. 1122–1128, 2012.

[11] C. Zhou *et al.*, "Updated efficacy and quality-of-life (QoL) analyses in OPTIMAL, a phase III, randomized, open-label study of first-line erlotinib versus gemcitabine/carboplatin in patients with EGFR-activating mutation-positive (EGFR Act Mut+) advanced non-small cell lung cancer (NSCLC)," *Journal of Clinical Oncology*, vol. 29, no. 15_suppl, pp. 7520–7520, 2011.

[12] C. G. Azzoli *et al.*, "2011 focused update of 2009 American Society of Clinical Oncology clinical practice guideline update on chemotherapy for stage IV non–small-cell lung cancer," *Journal of clinical oncology*, vol. 29, no. 28, pp. 3825–3831, 2011.

[13] M. H. Cohen, G. A. Williams, R. Sridhara, G. Chen, and R. Pazdur, "FDA drug approval summary: gefitinib (ZD1839)(Iressa®) tablets," *The oncologist*, vol. 8, no. 4, pp. 303–306, 2003.

[14] G. Yin *et al.*, "Prediction of EGFR mutation status based on 18F-FDG PET/CT imaging using deep learning-based model in lung adenocarcinoma," *Frontiers in Oncology*, vol. 11, p. 709137, 2021.

[15] Y. Dong *et al.*, "Multi-channel multi-task deep learning for predicting EGFR and KRAS mutations of non-small cell lung cancer on CT images," *Quantitative imaging in medicine and surgery*, vol. 11, no. 6, p. 2354, 2021.

[16] Y. Liu *et al.*, "Radiomic features are associated with EGFR mutation status in lung adenocarcinomas," *Clinical lung cancer*, vol. 17, no. 5, pp. 441–448. e6 2016.

[17] E. Rios Velazquez *et al.*, "Somatic mutations drive distinct imaging phenotypes in lung cancer," *Cancer research*, vol. 77, no. 14, pp. 3922–3930 2017.

[18] M. Jafari, B. Ghavami, and V. Sattari Naeini, "The Predictor Set Inference in Gene Regulatory Network Using Gravitational Search Algorithm and Mean Conditional Entropy Fitness Measure," *Iranian*

Journal of Biomedical Engineering, vol. 9, no. 4, pp. 375–386, 2016.
[19] J.-P. Jiang, S.-Y. Liu, H.-R. Cai, Q. Zhou, and H.-J. Ye, "Representation learning for tabular data: A comprehensive survey," *arXiv preprint arXiv:2504.16109*, 2025.
[20] W.-Y. Wang, W.-W. Du, D. Xu, W. Wang, and W.-C. Peng, "A survey on self-supervised learning for non-sequential tabular data," *Machine Learning*, vol. 114, no. 1, p. 16, 2025.
[21] X. Huang, A. Khetan, M. Cvitkovic, and Z. Karnin, "Tabtransformer: Tabular data modeling using contextual embeddings," *arXiv preprint arXiv:2012.06678*, 2020.
[22] Y. Zheng *et al.*, "A scoping review of self-supervised representation learning for clinical decision making using EHR categorical data," *npj Digital Medicine*, vol. 8, no. 1, p. 362, 2025.
[23] X. Li *et al.*, "Knowledge enhanced representation learning network for drug recommendation," *Information Processing & Management*, vol. 62, no. 5, p. 104164, 2025.
[24] K. Dradjat, M. Hamidi, P. Bartet, and B. Hanczar, "Self-supervised representation learning on gene expression data," *Bioinformatics*, vol. 41, no. 11, p. btaf533, 2025.
[25] S. Bakr *et al.*, "A radiogenomic dataset of non-small cell lung cancer," *Scientific data*, vol. 5, no. 1, pp. 1–9 2018.
[26] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.
[27] S. Soleimani, M. Bahrami, and M. Vali, "Survival prediction from imbalanced colorectal cancer dataset using hybrid sampling methods and tree-based classifiers," *Scientific Reports*, vol. 15, no. 1, p. 14554, 2025.
[28] M. Bahrami, M. Vali, and H. Kia, "Breast Cancer Detection from Imbalanced Clinical Data: A Comparative Study of Sampling Methods," 2023: IEEE, pp. 145–149.
[29] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," 2016, pp. 785–794.
[30] T. Dozat, "Incorporating nesterov momentum into adam," 2016.