شناسایی سرطان سینه
با نمودار کنترل آماری چندمتغیره ناپارامتری

رباب افشاری [*۱]

۱ دانشگاه زنجان، دانشکده علوم، عضو هیات علمی گروه آمار، *afshari@znu.ac.ir*

* نویسنده مسئول

چکیده

سرطان سینه دومین سرطان رایج در دنیا بین زنان است که زندگی بسیاری از آنان را در ابعاد مختلف تحت تاثیر قرار می‌دهد. شناسایی و تشخیص زود هنگام این بیماری، کمک شایانی در معالجه و نجات مرگ ناشی از آن دارد. یکی از روش‌های آماری که اهمیت فزاینده‌ای در مطالعات نظارت حوزه پزشکی دارد، نمودارهای کنترل کیفیت آماری است. این نمودارها غالبا با فرض عدم خودهمبستگی مشاهدات و نرمال بودن توزیع آنها طراحی می‌شوند. اما در واقعیت، ممکن است فرض‌های مذکور برقرار نبوده و یا توریع داده‌ها نامعلوم باشد. در چنین شرایطی، نمودارهای کنترل مذکور از کارایی لازم برای نظارت برخوردار نخواهند بود. از این‌رو و برای رفع این مساله، در این مقاله نمودار کنترل کیفیت میانگین متحرک موزون نمایی ناپارامتری چند متغیره مبتنی بر توزیع آماره‌های ترتیبی ارائه می‌گردد. در ادامه برای تشریح چگونگی کاربرد روش معرفی‌شده در تشخیص و پیش‌بینی سرطان سینه، مثالی از داده‌های واقعی آورده می‌شود.

کلمات کلیدی: خود همبستگی، سرطان سینه، میانگین متحرک موزون نمایی، نمودار کنترل، آماره ترتیبی

۱- مقدمه

سرطان سینه، تومور بدخیم متشکل از مجموعه‌ای از سلول‌های سرطانی سینه است. اگرچه این نوع سرطان، عمدتا در میان زنان شایع است، اما در برخی مواقع می‌تواند مردان را نیز مبتلا سازد. یکی از استراتژی‌های مهم برای تشخیص زودهنگام سرطان سینه، غربالگری است که خود موجب می‌شود با تجویز دارو و درمان به موقع،

احتمال نتیجه درمانی خوب در بیمار افزایش یابد. تاکنون مدل‌های پیش‌بینی متنوعی مبتنی بر اطلاعات بدست-آمده از مشاوره‌های معمول و آنالیز خون بیماران توسط محققان ارائه شده‌اند که در کنار سایر ابزارهای غربالگری اهمیت بسیاری در تشخیص اینکه آیا تغییری در شاخص سرطان سینه وجود دارد یا خیر، دارد.

یکی از ابزارهای مهم آماری در بررسی و تشخیص عیب در عملکرد یک فرآیند تولیدی، نمودارهای کنترل کیفیت آماری است. یکی از اهداف اصلی این ابزارها این است که آیا تغییری در برآیند یا متوسط مشخصه کیفیت فرآیند از مقدار مطلوب آن ایجاد شده است یا نه، تا در صورت بروز تغییرات ناگهانی، اقدامات اصلاحی در زمان مناسب در فرآیند ایجاد شود. نمودارهای کنترل برای اولین بار توسط شوهارت برای مشخصه‌های کیفیت یک-متغیره مطرح شد [۱]. یکی از ضعف‌های عمده نمودارهای شوهارت عدم توانایی آنها در تشخیص تغییرات کوچک در میانگین فرآیند است. برای مرتفع ساختن این مساله، محققان طرح‌های جایگزین دیگری برای نظارت ارائه نمودند که از جمله آنها می‌توان به نمودارهای کنترل جمع تجمعی[۱] *(Cusum)* [۲] و میانگین متحرک موزون نمایی[۲] *(EWMA)* [۳] اشاره کرد که مناسب برای مشاهدات یک‌متغیره با توزیع نرمال است. هر نمودار کنترل دارای دو حد کنترل پایینی[۳] *(LCL)* و حد کنترل بالایی[۴] *(UCL)* است. در صورتی که مقدار آماره کنترل پس از ترسیم بین حدود کنترل قرار گیرد اصطلاحا گویند فرآیند تحت کنترل است و در غیر این صورت خارج از کنترل نامیده می‌شود. عملکرد هر نمودار کنترل بر اساس شاخص متوسط طول اجرای دنباله[۵] *(ARL)* ارزیابی می‌شود. منظور از ARL متوسط تعداد نقاطی است که داخل حدود کنترل قرار می‌گیرند قبل از اینکه یک نقطه خارج از کنترل مشاهده شود. متوسط طول اجرای دنباله برای فرآیند تحت کنترل (خارج از کنترل) را با ARL_0 (ARL_1) نشان می‌دهند. هرگاه یک فرآیند تولیدی تحت کنترل باشد، نمودار کنترلی دارای عملکرد خوب است که مقدار ARL_0 آن بزرگتر باشد یا به عبارت دیگر آلارم‌های تشخیص اشتباه در این نمودار دیرتر به صدا در آید. همچنین در حالتی که فرآیند خارج از کنترل باشد، ، نمودار کنترلی دارای عملکرد خوب است که مقدار ARL_1 آن کوچکتر باشد و یا به عبارت دیگر آلارم‌های تشخیص صحیح زودتر به صدا در آید.

از آنجا که در بسیاری از فرآیندها، کیفیت محصول می‌تواند متاثر از چندین مشخصه (متغیر) باشد، نمودارهای کنترل برای حالت چندمتغیره توسط محققان معرفی شده‌اند که از جمله آنها می‌توان نمودارهای کنترل T^2-هتلینگ، *Cusum* چند متغیره[۶] *(MCusum)* و *EWMA* چندمتغیره[۷] *(MEWMA)* را نام برد [۱].

امروزه نمودارهای کنترل علاوه بر صنایع تولیدی، در حوزه پزشکی نیز بسیار مورد توجه پژوهشگران قرار گرفته است. در منبع [۴] نویسندگان از نمودار کنترل *MCusum* برای تشخیص تغییرات در الگوهای فضایی

---

[1] *Cumulative sum*
[2] *Exponentially weighted moving average*
[3] *Lower control limit*
[4] *Upper control limit*
[5] *Average run length*
[6] *Multivariate Cusum*
[7] *Multivariate EWMA*

مربوط به داده‌های سرطان سینه در شمال شرقی ایالت متحده استفاده کرده‌اند. همچنین محققان در [۵] به منظور نظارت و تشخیص بهبود هر چند کوچک بیماری پس از جراحی قلب بیماران، از نمودار $MEWMA$ بهره گرفتند. برای مطالعه بیشتر درباره کاربرد نمودارهای کنترل در حوزه بهداشت و پزشکی می‌توان به [۶] و [۷] مراجعه کرد. شایان ذکر است در طراحی نمودارهای کنترل نامبرده دو فرض نرمال بودن داده‌ها و عدم وجود خودهمبستگی بین مشاهدات در نظر گرفته می‌شود. به عبارت دیگر در صورت عدم برقراری فرض‌های مذکور و یا در حالتی که توزیع داده‌ها نامعلوم باشد، نمودارهای کنترل فوق از مناسبت لازم در امر نظارت و تشخیص تغییر در فرآیند برخوردار نخواهند بود. به همین منظور لی و همکاران [۸]، نمودار کنترل $MEWMA$ در حالت ناپارامتری را پیشنهاد دادند. در این مقاله، با الهام از رویکرد لی و همکاران، ضمن معرفی نمودار کنترل $MEWMA$ مبتنی بر توزیع آماره‌های ترتیبی، مثالی کاربردی بر اساس داده‌های واقعی برای تشریح چگونگی کاربرد روش ارائه شده برای پیش‌بینی و تشخیص بیماری سرطان سینه آورده می‌شود.

۲- نمودار کنترل $MEWMA$ کلاسیک

چنانچه اشاره شد نمودار کنترل $MEWMA$، تعمیم‌یافته نمودار $EWMA$ در حالتی که مشخصه کیفیت چندمتغیره باشد، است. از جمله مزیت‌های این نمودار، حساسیت بالای آن در تشخیص تغییرات کوچک در میانگین فرآیند است زیرا در طراحی این نمودار علاوه بر بکارگیری اطلاعات کنونی فرآیند، از اطلاعات گذشته فرآیند نیز استفاده می‌شود. فرض کنید $\boldsymbol{X}_i = (X_{1,i}, X_{2,i}, \dots, X_{p,i})'$, $(i = 1,2, \dots, m)$ نمونه‌ای تصادفی از توزیع نرمال p متغیره $N_p(\boldsymbol{\mu}, \sum)$ باشد که در آن $\boldsymbol{\mu} = (\mu_1, \mu_2, \dots, \mu_p)'$ بردار میانگین و $\sum = [\sigma_{ij}],\ i,j = 1,2, \dots, p$ ماتریس واریانس کواریانس است که در آن σ_{ij} کواریانس بین متغیرهای iام و jام را نشان می‌دهد.

آماره $MEWMA$ را که به صورت زیر تعریف می‌شود، در نظر بگیرید:

$$\boldsymbol{Z}_i = \boldsymbol{R}\boldsymbol{X}_i + (\boldsymbol{I} - \boldsymbol{R})\boldsymbol{Z}_{i-1}, \qquad \boldsymbol{Z}_0 = (0,0, \dots, 0)' \tag{1}$$

که در آن $\boldsymbol{R}$ یک ماتریس قطری از مرتبه p متشکل از پارامترهای هموار کننده (مقادیر وزن‌دهی) به صورت $\boldsymbol{R} = diag(\gamma_1, \gamma_2, \dots, \gamma_k, \dots, \gamma_p)$، $0 < \gamma_k \leq 1$ و $\boldsymbol{I}$ یک ماتریس همانی از مرتبه p است. در صورتی که از نظر متخصصان هیچ اولویتی بین متغیرها نباشد معمولا $\gamma_1 = \gamma_2 = \cdots = \gamma_p = \gamma$ در نظر می‌گیرند. فرآیند خارج از کنترل نامیده می‌شود هرگاه $\boldsymbol{Z}_i' \sum_{\boldsymbol{Z}i}^{-1} \boldsymbol{Z}_i > H$ باشد که در آن $H = UCL > 0$ حد بالایی نمودار کنترل است و مقدار آن از طریق شبیه‌سازی برای تحقق مقدار پیش‌فرض برای نرخ آلارم اشتباهی ARL_0 بدست می‌آید [۹]. چون همواره مقدار عبارت $\boldsymbol{Z}_i' \sum_{\boldsymbol{Z}i}^{-1} \boldsymbol{Z}_i$ مثبت است از این‌رو حد پایین در این نمودار برابر $LCL = 0$ است. همچنین ماتریس واریانس کواریانس آماره $\boldsymbol{Z}_i$ یعنی $\sum_{\boldsymbol{Z}i}$ برابر است با:

$$\sum\nolimits_{Zi} = \frac{\gamma}{2-\gamma}[1 - (1-\gamma)^{2i}] \sum \quad . \tag{2}$$

در صورتی‌که $i \to \infty$، داریم:

$$\sum\nolimits_{zi} = \frac{\gamma}{2-\gamma} \sum \qquad (3)$$

لازم به ذکر است برای تشخیص تغییرات کوچک (بزرگ) در میانگین فرآیند، مقادیر کوچک (بزرگ) برای پارامتر هموارکننده γ در نظر گرفته می‌شود [۱].

۳- نمودار کنترل *MEWMA* ناپارامتری مبتنی بر آماره‌های ترتیبی

چنانچه پیشتر اشاره شد در حالت کلاسیک دو شرط اصلی در طراحی نمودار *MEWMA*، نرمال بودن توزیع داده‌ها و عدم وجود خودهمبستگی بین مشاهدات است. در صورتی که این فرضیات برقرار نبوده و یا توزیع داده‌ها نامعلوم باشد، این روش از کارایی لازم در نظارت فرآیند برخوردار نخواهد بود. برای رفع این مساله و به منظور طراحی نمودار کنترل مناسب و با الهام از رویکرد [۸]، از مفهوم رتبه و بازنویسی آماره نمودار بر اساس آماره‌های ترتیبی بهره گرفته خواهد شد. متغیرهای مستقل $X_1, X_2, \dots$ از توزیع زیر را در نظر بگیرید:

$$X_i \sim \begin{cases} F(X, \mu_0), & i = 1,2, \dots, \tau \\ F(X, \mu_1), & i = \tau+1, \tau+2, \dots \end{cases} \qquad (4)$$

که در آن F یک توزیع پیوسته نامعلوم، μ_0 و μ_1 به ترتیب میانگین فرآیند در حالت تحت کنترل و خارج از کنترل و τ نقطه تغییر میانگین توزیع و نامعلوم است. فرض کنید R_i نشان‌دهنده رتبه X_i در بین مشاهدات $X_1, X_2, \dots, X_i, \dots, X_n$ باشد به طوری که $R_i = \sum_{j=1}^{i} I(X_i \geq X_j)$ است. که در آن تابع $I(.)$ نشان‌دهنده تابع نشانگر است.

در این صورت رتبه استاندارد شده ترتیبی برابر $R_i^* = (R_i - ER_i)/\sqrt{VarR_i}$ ، $(i \geq 2)$ است که در آن $ER_i = \frac{i+1}{2}$ ، $VarR_i = \frac{(i+1)(i-1)}{12}$ و R_i دارای توزیع یکنواخت $U(1,i)$ است. به سادگی می‌توان نشان داد وقتی $i \to \infty$ داریم: $R_i^* \sim U(-\sqrt{3}, \sqrt{3})$. حال m مشاهده مستقل از یک توزیع پیوسته چندمتغیره نامعلوم یعنی $\boldsymbol{X}_i = (X_{1,i}, X_{2,i}, \dots, X_{p,i})'$ ، $(i = 1,2, \dots, m)$ را در نظر بگیرید. فرض کنید $X_{j,1}, X_{j,2}, \dots, X_{j,m}$ ، m $(j = 1,2, \dots, p)$ مشاهده مربوط به jامین متغیر و $R_{j,i}^*$ نشان‌دهنده رتبه استاندارد شده jامین مولفه در iامین مشاهده یعنی $X_{j,i}$ باشد. در آن صورت بردار $\boldsymbol{Q}_i$ متشکل از $R_{j,i}^*$ ها را تشکیل می‌دهیم یعنی $\boldsymbol{Q}_i = (R_{1.i}^*, R_{2,i}^*, \dots, R_{p,i}^*)'$ که در آن هر مولفه $R_{j,i}^*$ هم‌توزیع با R_i^* است. در این صورت آماره *MEWMA* در رابطه (۱) بر اساس آماره ترتیبی (بردار رتبه $\boldsymbol{Q}_i$) به صورت زیر بازنویسی و آماره نمودار در حالت ناپارامتری به صورت زیر تعریف می‌شود:

$$\boldsymbol{W}_i = \boldsymbol{R}\boldsymbol{Q}_i + (\boldsymbol{I} - \boldsymbol{R})\boldsymbol{W}_{i-1}, \qquad \boldsymbol{W}_0 = (0,0,\dots,0)' \tag{5}$$

به طور مشابه با حالت کلاسیک، فرآیند در این حالت خارج از کنترل در نظر گرفته می‌شود هرگاه $T_i^2 = \boldsymbol{W}_i' \sum_{\boldsymbol{W}i}^{-1} \boldsymbol{W}_i > H'$ باشد که در آن H' حد کنترل بالایی نمودار است و مقدار آن با روشی مشابه با آنچه که در بخش قبل اشاره شد، محاسبه می‌شود. همچنین داریم:

$\sum_{Wi} = \frac{\gamma}{2-\gamma}\sum_Q$ که در آن $\sum_Q$ ماتریس واریانس کواریانس مربوط به بردار رتبه $\boldsymbol{Q}_i$ است که از روی نمونه محاسبه می‌شود.

۴- تحلیل داده‌های سرطان

در این مطالعه از مجموعه داده‌های سرطان سینه کویمبرا برگرفته از سایت مخزن داده یادگیری ماشین کویمبرا استفاده می‌شود [۱۰]. در این مجموعه داده نه ویژگی بالینی از ۱۱۶ شرکت‌کننده مشاهده و اندازه گیری شد. این مشخصات شامل اندازه‌های نه متغیر سن، *BMI*، گلوکز، انسولین، لپتین، آدیپنکتین، رزیستین و *MCP-1* است. طراحی هر نمودار کنترل شامل دو فاز است. در فاز اول، نمودار کنترل بر اساس مشاهداتی که در حالت تحت کنترل بودن فرآیند گرفته شده‌اند، ساخته می‌شود. سپس در ادامه کار (فاز دوم)، از نمودار ساخته شده در مرحله قبل برای نظارت و بررسی فرآیند در آینده استفاده می‌شود. در این مطالعه، پس از پیش‌پردازش داده‌ها شامل مدیریت داده‌های گمشده، هم‌مقیاس‌سازی، ... از اطلاعات مربوط به ۵۰ شرکت‌کننده سالم برای ایجاد نمودار کنترل در فاز اول استفاده شده است. در ادامه مراحل اجرای کار به ترتیب اجرا آورده می‌شود:

گام۱-اجرای آزمون وابستگی بین متغیرها: در صورتی که نتیجه آزمون مبین استقلال بین متغیرها باشد، در این صورت می‌توان از نه نمودار *EWMA* یک‌متغیره به جای نمودار *MEWMA* چند متغیره استفاده کرد.

گام۲-اجرای آزمون خودهمبستگی بین مشاهدات متناظر با هر متغیر: در صورت وجود خودهمبستگی معنادار بین مشاهدات، در طراحی نمودار از مقادیر باقی‌مانده‌ها به جای مشاهدات واقعی استفاده می‌شود [۱]. برای محاسبه باقی‌مانده‌ها می‌توان از برازش مدل به مشاهدات استفاده کرد که در این مطالعه از مدل رگرسیون غیرخطی براساس حداقل مربعات خطا استفاده شده است. لازم به ذکر است اگر $\hat{y}$ مقدار برآورد برای y بر اساس هر مدل برازش‌شده‌ای باشد آنگاه مقدار باقی‌مانده از رابطه $y - \hat{y}$ بدست می‌آید.

گام۳-اجرای آزمون نرمال بودن باقی‌مانده‌ها

گام۴-طراحی نمودار کنترل $MEWMA$ با اجرای مراحل زیر:

-در نظر گرفتن مقدار پیش‌فرض برای نرخ آلارم اشتباهی ARL_0

-محاسبه بردار $\boldsymbol{W_i}$ با استفاده از رابطه (۵) به ازای هر مشاهده و ماتریس واریانس کواریانس $\sum_{Wi}$

-محاسبه مقدار T_i^2 به ازای هر مشاهده

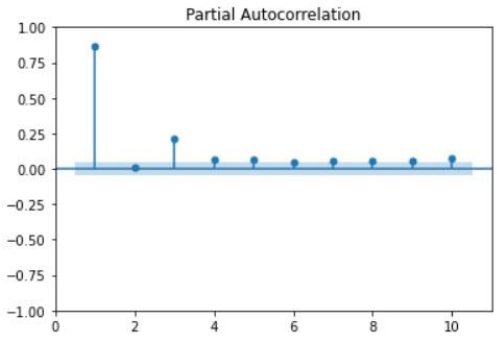
-محاسبه مقدار حد کنترل بالا $UCL = H'$ از طریق شبیه‌سازی و بر اساس مقادیر پیش‌فرض ARL_0 به ازای مقادیر مختلف γ

در این مطالعه از آزمون بارتلت[1] در اندازه $\alpha = 0.05$ برای بررسی اینکه آیا متغیرها از هم مستقل هستند یا خیر، استفاده شد. در این آزمون، مقدار احتمال $p.value = 0.01 < 0.05$ بدست آمد که مبین عدم استقلال متغیرها است. فلذا نمودار کنترل چندمتغیره انتخاب مناسبی برای بررسی تحت کنترل بودن یا نبودن فرآیند است. به منظور بررسی وجود یا عدم وجود خودهمبستگی بین مشاهدات هر متغیر، نمودار تابع[2] $PACF$ مربوط به هر کدام از متغیرها ترسیم می‌شود که نمودارهای مربوطه بیانگر وجود خودهمبستگی با تاخیر[3] یک هستند. به عنوان مثال، شکل ۱ نمودار $PACF$ متناظر با متغیر گلوکز را نشان می‌دهد. بنابراین به دلیل وجود خودهمبستگی بین مشاهدات از مقادیر باقی‌مانده‌ها به جای مشاهدات واقعی در طراحی نمودار استفاده می‌شود. نمودارهای مربوط به بقیه متغیرها به دلیل محدودیت در تعداد صفحات مقاله آورده نشده است.

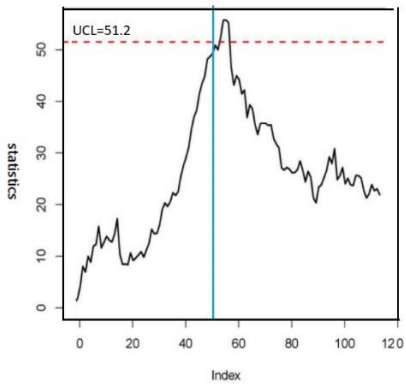
در مرحله بعد برای بررسی اینکه آیا باقی‌مانده‌ها از توزیع نرمال پیروی می‌کنند یا نه، از آزمون نرمالیتی چند متغیره خی دو استفاده شد. در این آزمون مقدار احتمال $p.value = 0.012 < 0.05$ است که فرض نرمال بودن داده‌ها را رد می‌کند. بنابراین نمودار کنترل $MEWMA$ ناپارامتری انتخاب مناسب برای نظارت و بررسی مجموعه داده‌های سرطان است. با فرض $ARL_0 = 370$ مقدار $UCL = H' = 51.2$ به ازای $\gamma = 0.3$ حاصل شد. شکل ۲، حدود کنترل نمودار $MEWMA$ ناپارامتری را به همراه مقادیر T_i^2 نشان می‌دهد. با توجه به شکل ۲ ملاحظه می‌شود مقدار آماره کنترل متناظر با مشاهده ۵۳ام در خارج از حدود کنترل قرار گرفته است و این بدان معنی است که شاخص‌های مربوط به بیماری سرطان در شرکت‌کننده $\#53$ خارج از محدوده حدود کنترل بوده و ریسک بالایی برای ابتلا به سرطان سینه دارد بنابراین باید مداخلات دارویی و درمانی برای او هر چه سریع‌تر آغاز گردد.

---

[1] *Bartlet*
[2] *Partial autocorrelation function*
[3] *Lag*



شکل ۱: نمودار خودهمبستگی جزئی مربوط به مشاهدات متغیر گلوکز



شکل ۲: نمودار کنترل *MEWMA* ناپارامتری برای مجموعه داده‌های سرطان سینه

۴- نتیجه‌گیری

سرطان سینه یکی از سرطان‌های رایج در دنیا است که شیوع آن در بین زنان نسبت به مردان بیشتر است. تشخیص زود هنگام سرطان سینه در اغلب موارد موجب درمان زودتر در مدت زمان کمتر می‌شود به طوری که هر چه این بیماری زودتر تشخیص داده شود شانس زنده ماندن بیمار افزایش می‌یابد. یکی از روش‌های تشخیص و پیش‌بینی این بیماری استفاده از نمودارهای کنترل آماری چند متغیره است. این نمودارها غالبا بر اساس فرض عدم وجود خودهمبستگی بین مشاهدات و نرمال بودن توزیع داده‌ها طراحی می‌شوند. اما از آنجا که در واقعیت ممکن است توزیع داده‌ها نرمال نبوده و یا نامعلوم باشد و از سوی دیگر خودهمبستگی معنادار بین مشاهدات وجود داشته باشد انتخاب این نمودارها در چنین شرایطی برای بررسی و نظارت بیماری، انتخاب مناسبی نخواهد بود. از این‌رو، در این مقاله نمودار کنترل میانگین متحرک موزون نمایی چند متغیره مبتنی بر توزیع آماره‌های ترتیبی با ملاحظه خودهمبستگی احتمالی بین مشاهدات ارائه شد. در پایان، برای تشریح چگونگی استفاده از روش ارائه شده، مثال کاربردی از مجموعه داده‌های واقعی مربوط به سرطان سینه آورده شد.

مراجع

[1] *D.C. Montgomery, (2020) "Introduction to Statistical Quality Control". Wiley, New York.*

[2] *R.B. Crosier, "A new two-sided cumulative sum quality control scheme", Technometrics. 1986;28(3):187–194.*

[3] *J.S. Hunter, "The exponentially weighted moving average", J. Qual. Technol., 1986;18(4):203–210.*

[4] *P.A. Rogerson, & I. Yamada, "Monitoring change in spatial patterns of disease: comparing univariate and multivariate cumulative sum approaches", Stat Med. 2004;23(14):2195–2214.*
[5] *M. Abdollahian & P. Hayati Rezvan, "Multivariate exponentially weighted moving average chart for monitoring patient's progress after cardiac surgery", in Proceedings of the 2012 World Congress in Computer Science-Computer Engineering and Applied Computing. Las Vegas, USA: 2012, 16–19.*
[6] *A. Yeganeh, A. Johannssen, N. Chukhrova & M. Rasouli, "Monitoring multistage healthcare processes using state space models and a machine learning based framework", Artificial Intelligence in Medicine, 2024, 151, 102826.*
[7] *R. Sasikumar, (2023) "Applications of Statistical Quality Control Charts in Public Health and Epidemiology". Shashwat Publication.*
[8] *L. Liu, X. Zi, J. Zhang & Z. Wang, "A sequential rank–based nonparametric adaptive EWMA control chart", Commun Stat Simul Comput, 2013,42(4),841–859.*
[9] *S.S. Prabhu & G.C. Runger, "Designing a multivariate EWMA control chart", J. Qual. Technol., 1997, 29(1), 8.*
[10] *http://archive.ics.uci.edu/mL/datasets/Breast+Cancer+Coimbra.*