Visualizing Cancer Survivability Using R Language

Tarikul Islam Nishat
CSE
American International University - Bangladesh (AIUB)
21-44632-1

Mohammad Rafiul
CSE
American International University - Bangladesh (AIUB)
21-44631-1

Sadaf Akhter
CSE
American International University - Bangladesh (AIUB)
21-44658-1

Md.Sanim
CSE
American International University - Bangladesh (AIUB)
21-44606-1

ABSTRACT: *Cancer's insidious nature often evades detection, profoundly impacting individuals and communities alike. Addressing this challenge requires innovative approaches to illuminate key insights into cancer dynamics. Leveraging extensive datasets encompassing cancer incidence, patient demographics, and treatment outcomes, this paper aims to unveil intricate patterns in cancer epidemiology, including regional disparities and temporal trends. By employing insightful data visualization techniques, we enhance transparency and comprehension of complex datasets, providing visually accessible insights. These visualizations foster a deeper understanding of underlying data distributions, facilitating optimized treatment regimens and ultimately improving survivability rates and quality of life. Through advancing early detection initiatives, refining treatment protocols, and informing evidence-based policy decisions, ourr approach aims to catalyze advancements in cancer research and clinical practice, while providing a deeper understanding of data visualization and its techniques using the R language*.

KEYWORDS:
Data, Data visualization, Cancer, Visualization techniques

I Introduction

In today's era dominated by data, the sheer amount and diversity of information produced across numerous sectors and fields are unparalleled.. Data serves as the foundation upon which critical decisions are made, insights are obtained, and innovations are driven forward. However, the mere existence of data is not enough; its significance lies in the actionable insights it harbors; while data holds immense potential, its true value is unlocked through effective visualization. This is where data visualization emerges as an indispensable tool, amplifying the value and impact of data in numerous ways. Data visualization serves as a bridge between raw data and human comprehension. While data in its raw form may be dense, complex, and difficult to interpret, visual representations transform it into intuitive, understandable formats. Visualizations leverage human visual perception, enabling individuals to identify patterns, trends, and outliers more effectively than through raw data alone. By presenting data in visually appealing and digestible formats, data visualization empowers stakeholders to extract meaningful insights swiftly and accurately by transforming raw data into compelling visual narratives. Visualization enhances understanding, communication, and exploration, empowering individuals and organizations to make informed decisions, drive innovation, and create positive change. In today's era abundant with data, the significance of data visualization cannot be emphasized enough. It serves as the gateway to unleashing the profound potential held within data..

This paper seeks to offer a thorough examination of cancer survival rates, alongside an introduction to utilizing data visualization methods, particularly focusing on recent developments across various fields. It will utilize the R language to effectively visualize pertinent insights derived from a Cancer Survivability Dataset. We examined eight notable contributions from academic literature, journals and conferences, each offering unique insights into the evolving landscape of data visualization.

II Literature Review

In data visualization, there are significant advancements, with researchers exploring innovative techniques and tools to represent and interpret data effectively across various domains. This literature review examines eight notable contributions that offer insights into different aspects of data visualization.

The paper in [1] provides a broad overview of data visualization techniques within the context of engineering research and advanced technology. The paper discusses

various visualization techniques utilized in engineering contexts, including; line graph: which shows the relationship between items comparing changes over a period of time, bar chart: which is used to compare quantities of different categories, scatter plot: this is a two-dimensional plot showing variation of two items, pie chart: which is used to compare the parts of a whole.

In [2], the author focuses on introducing data visualization tools and techniques across various domains, rather than a specific field like engineering. The paper offers a comprehensive overview of popular tools and techniques used for data visualization, spanning domains such as business, healthcare, social sciences, and more. The author discusses a range of tools, both traditional and emerging, including but not limited to: business intelligence (BI) tools like Tableau, Power BI, and QlikView, programming languages and libraries such as R, Python (matplotlib, seaborn), and JavaScript (D3.js), visualization software like Microsoft Excel and Google Data Studio, specialized tools for specific domains, such as medical imaging software in healthcare or GIS tools for geographic data analysis. Moreover, the author discusses upon various visualization techniques employed in different domains, such as interactive dashboards, infographics, heatmaps, and more. By providing a comprehensive survey of data visualization tools and techniques across diverse domains, the paper ikely serves as a valuable resource for researchers, practitioners, and students looking to explore and utilize In the era of big data, the ability to effectively communicate insights derived from complex datasets is paramount across various disciplines, ranging from engineering and psychology to statistics and beyond. Data visualization, the process of representing data visually to facilitate understanding, has emerged as a crucial tool in this endeavor. Advancements in data visualization techniques not only enable researchers and practitioners to explore and analyze data more efficiently but also enhance communication and decision-making processes.

This paper aims to provide a comprehensive overview and analysis of recent advancements in data visualization techniques across different domains. We examine eight notable contributions from academic literature, spanning journals, conferences, and electronic publications, each offering unique insights into the evolving landscape of data visualization.

visualization methods in their respective fields.

While both [1] and [2] offer valuable insights into data visualization tools and techniques, they cater to slightly different audiences and focus areas. While the [1] delves into visualization within the specific context of engineering and advanced technology, [2] offers a broader perspective encompassing various domains. Together, these contributions contribute to the ongoing discourse surrounding the exploration and utilization of data visualization tools and techniques across different fields, ultimately advancing the practice and application of visualization in research and industry.

In [3], the authors primarily focus on leveraging the capabilities of the R programming language for data visualization. They utilized a variety of R packages and libraries commonly used for visualization tasks, such as ggplot2, which is renowned for its flexibility and ease of use in creating high-quality graphics. In terms of methods, the authors demonstrated a range of visualization techniques using R, including but not limited to; scatter plots: for visualizing relationships between variables, bar charts for comparing categorical data, line graphs: for displaying trends over time or continuous variables, heatmaps: for illustrating patterns or correlations in large datasets, boxplots: for summarizing distributions and identifying outliers, violin plots: for combining aspects of boxplots and kernel density plots to visualize distributions.

On the other hand, the authors in [4] offer broader guidance on data visualization principles rather than focusing on specific tools or programming languages. They may not explicitly mention specific libraries or methods, but their recommendations likely encompass general principles applicable across various visualization platforms. These principles include: emphasizing clarity and simplicity in visualization design, ensuring accuracy and integrity in representing data, incorporating principles of visual perception and cognition to enhance comprehension, using color effectively to convey information and facilitate interpretation, avoiding chart junk and unnecessary embellishments that detract from the message, considering the audience and context when designing visualizations.

While [4] paper may not delve into the technical details of implementing specific visualization methods, their emphasis on fundamental principles provides a valuable framework for researchers to apply regardless of the tools or libraries they use.

In [5], the authors present a novel approach titled "An Third Eye of Data Visualization Using R," published in the SSRN Electronic Journal. By framing data visualization as a "third eye," the authors propose R-based methods or techniques that enable researchers to gain deeper insights into their data, potentially through advanced statistical analyses, interactive visualizations, or integration with other data analysis tools.

This paper contributes to the ongoing exploration of R's capabilities as a versatile platform for creating insightful and impactful visualizations, catering to researchers and practitioners seeking to leverage R's analytical power in their visualization workflows.

In [7], the authors explore the intersection of machine learning (ML) and data visualization, recognizing the potential for ML techniques to enhance the creation, analysis, and interpretation of visual representations of data. Through their survey, the author likely review a wide range of ML-driven approaches to data visualization, including but not limited to; automated visualization generation: ML algorithms that automatically generate visualizations based on data characteristics and user preferences, data-driven visualization recommendations: ML models that suggest visualization types or design choices based on the underlying data structure and analysis goals, interactive and adaptive visualizations: ML-powered systems that enable real-time interaction and adaptation of visualizations in response to user input or changing data conditions, visual analytics with ML integration: Techniques that combine ML algorithms with traditional visualization methods to facilitate exploratory data analysis, pattern discovery, and hypothesis generation. By surveying the landscape of ML-driven advancements in data visualization, the author provides valuable insights into emerging trends, challenges, and opportunities at the intersection of machine learning and visualization. This paper serves as a foundational resource for researchers, practitioners, and enthusiasts interested in harnessing ML techniques to push the boundaries of data visualization capabilities, ultimately leading to more effective and insightful visual representations of complex datasets.

While [5] explores innovative approaches within the context of R-based data visualization, [7] offers a broader perspective on the integration of machine learning techniques into the data visualization process. Together, these contributions highlight the diverse avenues for innovation in data visualization techniques, paving the way for future advancements that promise to transform how we explore, understand, and communicate insights from data.

The paper in [8] focuses on addressing the challenges of visualizing large-scale datasets by leveraging the capabilities of two prominent programming languages, R and Python, supplemented with graphical user interface (GUI) tools. By combining the analytical power of R and Python with user-friendly GUI interfaces, the authors propose a practical approach to enable researchers and practitioners to explore and visualize big data effectively. This paper contributes to the ongoing discourse on big data visualization by providing insights into methodologies and tools that facilitate the analysis and interpretation of large and complex datasets, ultimately enabling stakeholders to derive actionable insights from big data resources.

In contrast, the authors in [9] offer a perspective on integrating data visualization techniques into introductory statistics education. The authors advocate for the early and frequent introduction of data visualization concepts and tools in statistics education curricula to foster a deeper understanding of statistical principles among students. By emphasizing the importance of visualization in conveying key statistical concepts and fostering data literacy skills, the authors propose strategies for incorporating data visualization exercises and activities into introductory statistics courses. This paper serves as a foundational resource for educators seeking to enhance statistical education by integrating data visualization components, ultimately empowering students to become proficient in interpreting and communicating insights from data.

While the paper in [8] addresses the challenges of visualizing big data using R, Python, and GUI tools, the author's work in [9] advocates for the integration of data visualization into statistics education to promote data literacy among students. Together, these contributions highlight the diverse applications and implications of big data visualization across research, education, and practice, underscoring its significance in extracting value from large and complex datasets.

III Methodology

The we conducted data analysis using a dataset obtained from the open-data portal, www.kaggle.com [6]. This dataset comprises 569 records and offers a diverse and extensive range of information pertaining to cancer research. It includes detailed data on cancer incidence, patient demographics, treatment outcomes, genomic profiles, and additional relevant variables. Notably, the dataset encompasses the following columns:

- id: Unique identifier for each tumor sample.
- diagnosis: The diagnosis of the tumor (M = malignant, B = benign).
- radius_mean: Mean radius of the tumor.
- texture_mean: Mean texture value of the tumor.

- perimeter_mean: Mean perimeter of the tumor.
- area_mean: Mean area of the tumor.
- smoothness_mean: Mean smoothness of the tumor.
- compactness_mean: Mean compactness of the tumor.
- concavity_mean: Mean concavity of the tumor.
- concave points_mean: Mean number of concave portions of the contour.
- symmetry_mean: Mean symmetry of the tumor.
- fractal_dimension_mean: Mean fractal dimension of the tumor.
- radius_se: Standard error of the radius of the tumor.
- texture_se: Standard error of the texture value of the tumor.
- perimeter_se: Standard error of the perimeter of the tumor.
- area_se: Standard error of the area of the tumor.
- smoothness_se: Standard error of the smoothness of the tumor.
- compactness_se: Standard error of the compactness of the tumor.
- concavity_se: Standard error of the concavity of the tumor.
- concave points_se: Standard error of the number of concave portions of the contour.
- symmetry_se: Standard error of the symmetry of the tumor.
- fractal_dimension_se: Standard error of the fractal dimension of the tumor.
- radius_worst: Worst (largest) radius of the tumor.
- texture_worst: Worst (largest) texture value of the tumor.
- perimeter_worst: Worst (largest) perimeter of the tumor.
- area_worst: Worst (largest) area of the tumor.
- smoothness_worst: Worst (largest) smoothness of the tumor.
- compactness_worst: Worst (largest) compactness of the tumor.
- concavity_worst: Worst (largest) concavity of the tumor.
- concave points_worst: Worst (largest) number of concave portions of the contour.
- symmetry_worst: Worst (largest) symmetry of the tumor.
- fractal_dimension_worst: Worst (largest) fractal dimension of the tumor.

The "diagnosis" column serves as a pivotal categorical variable indicating whether a patient's diagnosis is malignant ('M') or benign ('B'). This distinction provides crucial information for further analysis and classification tasks. After loading the dataset, we checked for missing values and then discarded those instances. Next, we checked for outliers by calculating values of the 1^{st} and 5^{th} whiskers in the boxplot. This helped us determine the valid range of value. We then discarded the instances with values outside the valid range. After this, we considered the dataset as a clean dataset and moved on to visualizing it. For visualizing the dataset, we employed a myriad of techniques such as heat map, histogram, area chart, scatter plot, violin plot, radar chart. The meaningful information extracted from visualizing the dataset using these techniques is further explained in the Results section.

IV Results

Heatmap:

A heatmap helps us distinguish patterns and trends in data by visualizing the relationship between two categorical variables or between one categorical and one continuous variable. It displays data as a grid of colored cells, where the color intensity represents the value of the variable being measured. Heatmaps are particularly useful for identifying correlations, clusters, and anomalies in large datasets. They provide a visual summary of the data distribution, highlighting areas of high or low concentration. Heatmaps are commonly used in fields such as data analysis, biology, finance, and geography to uncover insights and make data-driven decisions.

As shown in figure-1 the heatmap generated from these columns provides a visual representation of the correlations between various tumor characteristics. By examining the intensity of colors in the heatmap, we can identify both positive and negative correlations between different features. For instance, darker shades indicate stronger correlations, while lighter shades suggest weaker or negligible associations. This analysis offers valuable insights into the relationships among tumor attributes, potentially revealing underlying patterns or dependencies that can inform diagnostic and treatment strategies in cancer care.

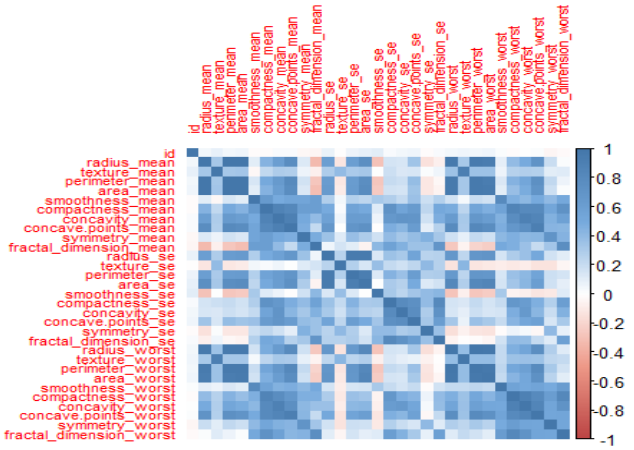

Figure-1: Heat map

Histogram:

Histograms help us distinguish the distribution and frequency of values within a dataset. By displaying the distribution of numerical data through bars, histograms provide insights into the central tendency, spread, and shape of the data.

From Figure-2, the histogram of smoothness in cancer diagnosis elucidates the distribution of smoothness values within the dataset, providing crucial insights into potential associations with cancer diagnosis. By examining the histogram, distinct patterns emerge, offering indications of tumor characteristics. A shift towards (right) higher smoothness values may suggest a higher incidence of benign tumors, while a distribution skewed towards (left) lower values might imply a prevalence of malignant tumors. Conversely, similar observations can be extrapolated for Figures 3 and 4. This analysis aids in understanding tumor behavior, aiding clinicians in diagnosis and treatment decisions for improved patient outcomes.

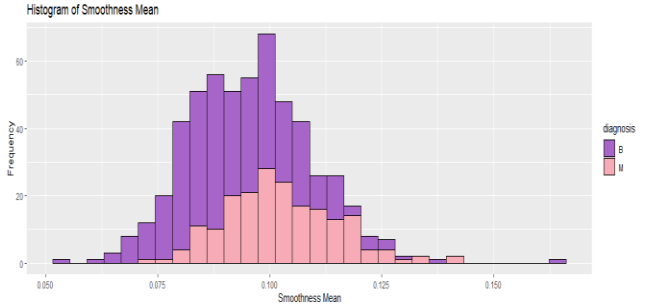

Figure-2: Histogram Smoothness mean

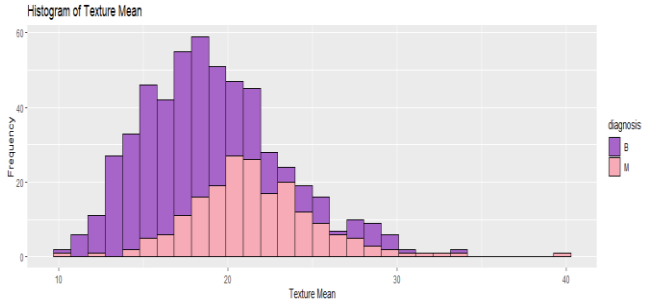

Figure-3: Histogram of Texture mean

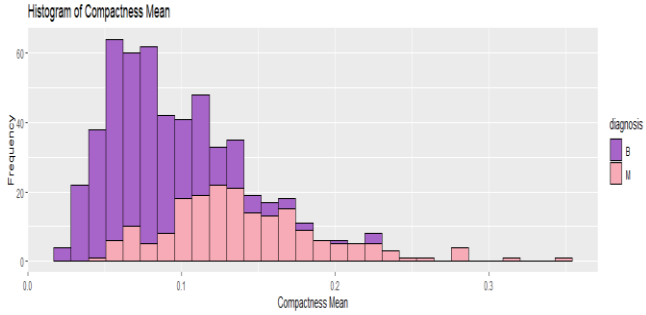

Figure-4: Histogram Compactness mean

Area chart:

Area charts help us distinguish trends and patterns in data over time or across categories. By displaying the magnitude of values as areas under lines connecting data points, area charts allow us to visualize changes and fluctuations in data more intuitively. This visualization technique enables us to identify trends, spot anomalies, and compare multiple datasets simultaneously. Additionally, area charts are effective for showcasing cumulative data and highlighting variations in the distribution of values.

The area chart in Figure - 5 depicting concavity and compactness in cancer diagnosis reveals distinct trends in these features among patients, offering potential associations with cancer diagnosis. Upon analysis, fluctuations or patterns in concavity and compactness values may indicate

variations in tumor characteristics. For instance, an upward trend in concavity levels may suggest a higher prevalence of malignant tumors with irregular cell contours, while fluctuations in compactness could signify changes in cell density or clustering.

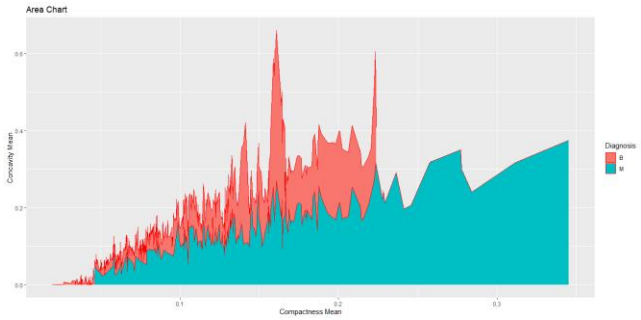

Figure-5: Area chart of Concavity mean, and Compactness mean

Scatter plot:

A scatter plot helps us distinguish the relationship between two variables. By plotting the data points and fitting a regression line through them, we can visualize the overall trend or pattern in the data. This type of plot helps us identify the direction and strength of the relationship between the variables, whether it's positive, negative, or no correlation. Additionally, the regression line can serve as a predictive tool, allowing us to estimate the value of one variable based on the value of the other. Overall, a scatter plot with a regression line provides valuable insights into the association between two variables, aiding in data analysis and interpretation.

From the scatter plot generated using ggplot shown in Figure – 6, it depicts the radius mean in cancer diagnosis patients and provides a visual representation of the relationship between tumor size and cancer diagnosis. Upon examination, distinct clusters or patterns may emerge, offering insights into the potential association between radius mean and malignancy. For instance, a separation between clusters of data points may indicate a correlation between larger tumor sizes and a higher likelihood of malignancy. Conversely, overlapping clusters could suggest less definitive conclusions about the relationship between tumor size and cancer classification.

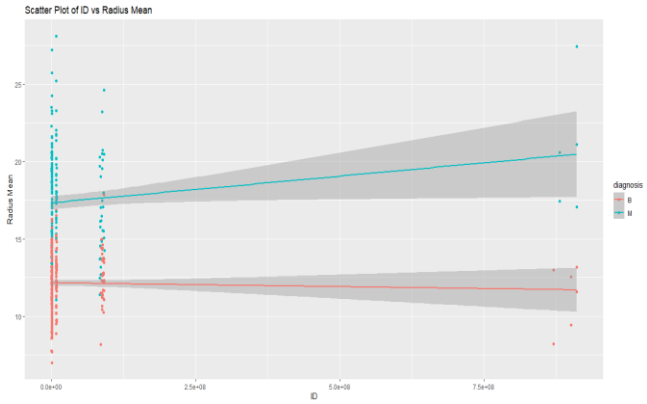

Figure-6: Scatter plot of Radius mean

In Figure - 7 we observe 'id' against 'radius_worst' with hue differentiation by 'diagnosis', a scatter of data points representing individual tumor samples. The linear regression line for each diagnosis group aids in visualizing any potential correlation between the tumor's worst radius and its identification as malignant or benign. Clusters or patterns in the data points, along with the trend lines, may suggest a relationship between tumor size and diagnosis. This visualization facilitates the identification of any discernible trends or outliers. Conversely, similar observations can be extrapolated for Figure - 8, the scatter plot with 'id' on the x-axis and 'smoothness_mean' on the y-axis.

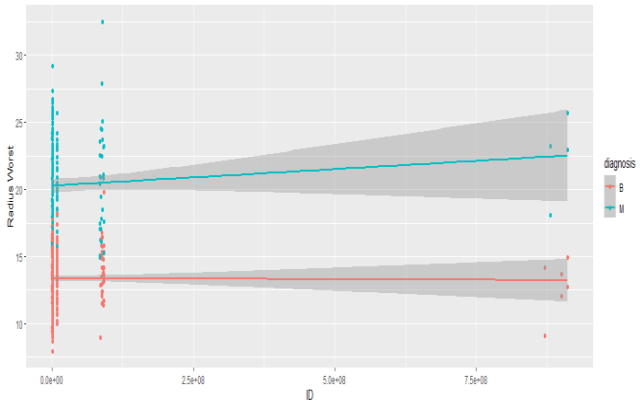

Figure-7: Scatter plot of Radius_worst

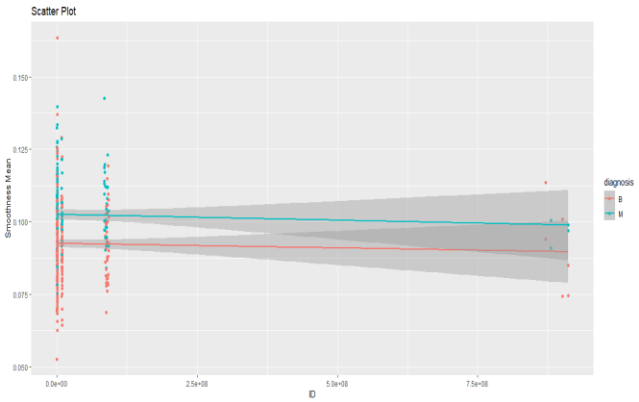

Figure-8: Scatter plot of Smoothness mean

Violin plot:

A violin plot helps us distinguish the distribution of data between different categories or groups. It provides a visual representation of the probability density of the data at different values, allowing us to compare the distribution shapes and spread of multiple groups simultaneously. This can be particularly useful for identifying differences in the central tendency, variability, and overall patterns between groups.

From the plot in Figure-9, we observe the comparison of mean values between malignant (M) and benign (B) diagnoses across several attributes: smoothness, compactness, concavity, concave points, and symmetry. In general, malignant tumors tend to have higher values across these attributes compared to benign tumors. Specifically, for compactness, concavity, and concave points, the distribution for malignant tumors is wider and shifted towards higher values, indicating greater irregularity and aggressiveness in malignant tumors. Conversely, benign tumors exhibit lower mean values and a narrower distribution for these attributes, suggesting a more uniform and less aggressive nature. The violin plots and boxplots visually depict these differences, highlighting the distinct characteristics between malignant and benign tumors in terms of these mean attributes.

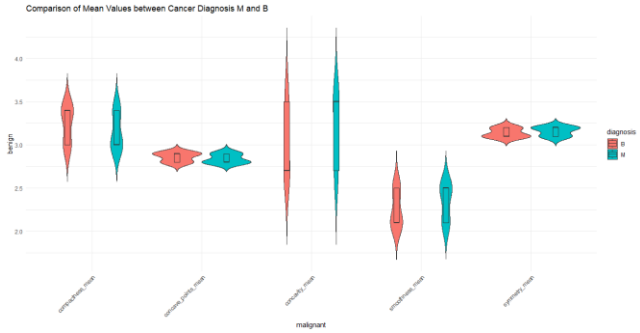

Figure-9: Violin chart of perimeter worst

Radar Chart:

Radar charts help us distinguish and compare multiple variables across different categories or groups. They provide a visual representation of how each variable compares relative to others within each category, allowing for easy identification of patterns, trends, and differences.

From the radar chart shown in Figure-10 compares the charactistics of cancer patients categorized as M (Malignant) in red color and B (Benign) in blue color, we observe distinct patterns in various attributes:

- Malignant patients generally exhibit higher values for attributes such as smoothness mean, concavity mean, and texture compared to benign patients.
- Conversely, benign patients tend to have higher values for smoothness_worst, concavity_worst, compactness_mean, concave_points_mean, and symmetry_mean.
- This indicates that there are notable differences in the characteristics of cancer patients based on their diagnosis (Malignant or Benign), suggesting potential markers for distinguishing between the two categories.

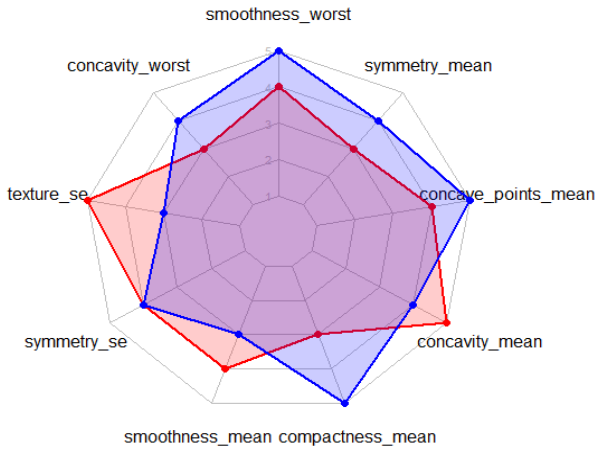

Figure-10, Malignant attributes are depicted in red, while Benign attributes are represented in blue on the radar chart.

IV Conclusion and Future Scope

In our data analysis, we delved into various interrelations among different parameters, comparing our findings between individuals diagnosed with cancer and those without. Such analyses are invaluable, shedding light on potential symptom correlations and aiding in predictive modeling to preemptively mitigate adverse health outcomes. As healthcare data analysis continues to surge forward, researchers are actively exploring new avenues to dissect and comprehend complex datasets. However, challenges such as data incompleteness and veracity persist. Nevertheless, innovative algorithms are being developed to confront these hurdles head-on. Looking ahead, the future holds promise for even more advanced techniques to unravel the intricacies of healthcare data, offering unprecedented insights and opportunities for transformative advancements in patient care and disease management.

REFERENCES

[1] "DATA VISUALIZATION | International Journal of Engineering Research and Advanced Technology (ijerat)," ijerat.com, Available: https://ijerat.com/index.php/ijerat/article/view/191

[2] D. Srivastava, "An Introduction to Data Visualization Tools and Techniques in Various Domains," International Journal of Computer Trends and Technology, vol. 71, no. 4, pp. 125–130, Apr. 2023, doi: https://doi.org/10.14445/22312803/ijctt-v71i4p116.

[3] E. Nordmann, P. McAleer, W. Toivo, H. Paterson, and L. M. DeBruine, "Data Visualization Using R for Researchers Who Do Not Use R," Advances in Methods and Practices in Psychological Science, vol. 5, no. 2, p. 251524592210746, Apr. 2022, doi: https://doi.org/10.1177/25152459221074654.

[4] E. Hehman and S. Y. Xie, "Doing Better Data Visualization," Advances in Methods and Practices in Psychological Science, vol. 4, no. 4, p. 251524592110453, Oct. 2021, doi: https://doi.org/10.1177/25152459211045334.

[5] H. R. Lekkala and V. Maddineni, "An Third Eye of Data Visualization Using R," SSRN Electronic Journal, 2020, doi: https://doi.org/10.2139/ssrn.3654792.

[6] https://www.kaggle.com/datasets/sophiyakanjirakadan/cancercsv/data

[7] Q. Wang, Z. Chen, Y. Wang, and H. Qu, "A Survey on ML4VIS: Applying Machine Learning Advances to Data Visualization," IEEE Transactions on Visualization and Computer Graphics, pp. 1–1, 2021, doi: https://doi.org/10.1109/tvcg.2021.3106142.

[8] S. A. Fahad and A. E. Yahya, "Big Data Visualization: Allotting by R and Python with GUI Tools," 2018 International Conference on Smart Computing and Electronic Enterprise (ICSCEE), Jul. 2018, doi: https://doi.org/10.1109/icscee.2018.8538413.

[9] X. Wang, C. Rush, and N. J. Horton, "Data Visualization on Day One: Bringing Big Ideas into Intro Stats Early and Often," Technology Innovations in Statistics Education, vol. 10, no. 1, 2017, doi: https://doi.org/10.5070/t5101031737.