A Review of Visualization Techniques for Duplicate Detection in Cancer Datasets

Nurul A. Emran[1]*, Ruhaila Maskat[2]
Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka, 76100, Melaka, Malaysia [1]
Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450, Selangor, Malaysia[2]

Abstract*—As clinical cancer research increasingly depends on large, diverse datasets, concerns about data duplication have grown. Duplicates can undermine data integrity, skew analytical results, and reduce the reproducibility of studies. This review explores how visualization can play a critical role in identifying and managing duplicates in non-image clinical cancer data. Drawing from literature in biomedical informatics, data quality, and visual analytics, it synthesizes current approaches and highlights key challenges. Using a scoping review methodology, we analyzed studies published over the past two decades, focusing on non-image clinical datasets. Studies were selected based on relevance to duplicate detection and visualization, excluding those centered on image or video data. Major datasets like The Cancer Genome Atlas (TCGA), The Cancer Imaging Archive (TCIA), and the North American Association of Central Cancer Registries (NAACCR) are examined to show how duplication occurs across genomic, clinical, and registry data. The review assesses existing visualization techniques based on their scalability, interactivity, integration with deduplication algorithms, and how well they address core data quality dimensions. While some tools offer scalable and interactive features, few provide clear visual representations of duplicates, especially those involving complex temporal and multidimensional patterns. Several methodological gaps are identified, including limited integration of data quality metrics, inadequate support for tracking changes over time, and a lack of standardized evaluation frameworks. To address these issues, the review advocates for the development of practical, user-friendly visualization tools that combine duplicate detection with key indicators of data quality. By offering a more complete and intuitive view of clinical datasets, such tools can help researchers and clinicians make better-informed decisions, ultimately improving the reliability and impact of cancer research. Bridging the gap between technical detection and visual understanding is essential for advancing data-driven healthcare and ensuring high-quality, reproducible outcomes.

Keywords*—Duplicate detection; data duplication; visualization; deduplication; TCGA; TCIA; NAACCR

I. Introduction

With the growing volume and complexity of biomedical data, especially in clinical cancer research, the issue of duplicate records has become increasingly important. Ensuring high data quality is essential, as duplication can significantly affect research accuracy and outcomes. As a result, visualizing duplicates in these datasets has emerged as a crucial area of study [1], [2]. In the past decade, improvements in data collection and integration have produced large, diverse clinical datasets. These datasets demand advanced visualization techniques to effectively assess data quality and support informed clinical decision-making [3], [4], [5], [6]. Duplicate records, where the same real-world entity appears multiple times can create serious challenges in data analysis. They may artificially boost predictive performance and lead to misleading research conclusions, ultimately compromising the reliability of clinical insights [1], [7]. Numerous studies have shown that duplicate records are common across various biomedical databases, including those containing transcriptomic and nucleotide sequence data. This highlights an urgent need for more effective methods to detect and visualize duplicates, ensuring data integrity and improving research outcomes [7]. Studies have shown that duplicate records can account for up to 30–50% of entries in some clinical databases, leading to inflated sample sizes, biased statistical outcomes, and compromised patient safety [8], [9]. The practical impact of duplicate records is clear as studies show they can lead to redundant or conflicting results, which in turn can disrupt downstream analyses and misguide clinical interpretations [7], [10]. This review focuses on a specific yet underexplored challenge: visualizing duplicate records in non-image clinical cancer datasets. Despite its importance, current methods in this area remain fragmented and underdeveloped, limiting their effectiveness in real-world clinical applications [1], [11].

Although many algorithms exist for detecting duplicates and assessing data quality, there's still a noticeable gap when it comes to visualization approaches specifically designed for the unique complexities of clinical cancer data. Current solutions often fall short in providing comprehensive, intuitive visual tools that support meaningful analysis in this domain [12], [13], [14]. There are differing views on the best approaches for identifying and representing duplicates, ranging from statistical record linkage methods to machine learning-based similarity measures. However, few studies have successfully combined these techniques with visualization tools that support human-in-the-loop data quality assessment, limiting their practical usefulness in clinical settings [15]–[17]. This gap can lead to serious consequences, including misinterpreting data, wasting time on inefficient curation efforts, and drawing flawed clinical insights that may affect patient care and research outcomes [7], [18]. Given the wide variety of data types and the complexity of clinical workflows, there's a clear need for visualization tools that can effectively handle multidimensional, temporal, and categorical data. These tools must be capable of adapting to the intricate nature of clinical cancer datasets to support accurate analysis and decision-making [19], [20].

This review introduces a conceptual framework that defines duplicates as multiple records referring to the same clinical

*Corresponding Author.

entity. It connects this definition to key data quality dimensions, such as completeness, accuracy, and consistency to emphasize the broader impact of duplication on clinical data integrity [12], [21]. Visualization is framed as a guided, interactive process that blends computational techniques for detecting duplicates with human pattern recognition. This approach enables users to actively assess and manage data quality, making the process more intuitive and effective [12], [22].

The framework brings together theories of data quality assessment and principles of visualization system design, highlighting how visual analytics can play a key role in managing clinical oncology data more effectively [3], [23]. This review aims to examine how current visualization tools and techniques are being applied to detect and represent duplicate records in clinical cancer datasets particularly those that do not involve image or video data. By drawing insights from fields such as biomedical informatics, data quality, and visual analytics, the review seeks to identify gaps in existing approaches and highlight opportunities for advancement.

A key contribution of this work is its effort to bridge the divide between technical duplicate detection algorithms and practical visualization tools that help researchers and clinicians interpret and manage data more effectively. This integration supports improved data curation and enhances the reliability of cancer research outcomes. To guide this exploration, the review adopts a scoping methodology, analyzing studies published over the past two decades. Literature searches were conducted across major academic databases, including PubMed, IEEE Xplore, Scopus, and ACM Digital Library, to ensure comprehensive coverage of relevant work.

The review focuses specifically on research related to duplicate detection and visualization in clinical or biomedical datasets, excluding studies centered on image or video data. This review deliberately excludes image and video data to focus on structured and semi-structured clinical datasets, such as EHRs, genomic records, and registries, where duplication manifests differently and is often harder to visualize. While this boundary narrows the scope, it allows for a more focused analysis of underexplored challenges in non-visual data domains, which are critical for data-driven decision-making in clinical workflows. The findings are organized to first outline the types of visualization techniques currently in use, then discuss the challenges and available tools, and finally propose directions for future research.

The remainder of the paper is organized as follows: Section II provides an overview of visualization approaches used in duplicate detection within clinical cancer datasets, focusing on technique diversity, data quality coverage, scalability, and user interaction. Section III presents a critical evaluation of methodological strengths and limitations, emphasizing integration with deduplication algorithms and temporal visualization. Section IV discusses the theoretical and practical implications of duplicate visualization, emphasizing the need for adaptive, context-aware frameworks and scalable visual tools. It highlights the integration of temporal stability and multidimensional analytics, advocating for human-in-the-loop systems that enhance data quality, support clinical decision-making, and improve research reliability. Section V introduces key datasets relevant to duplicate detection and visualization, discussing their characteristics and challenges. Section VI outlines existing gaps and proposes future research directions to advance the field. Finally, Section VII concludes the paper by summarizing key findings and highlighting the importance of integrating visualization with data quality assessment to improve clinical research outcomes.

II. Overview of Visualization Approaches in Duplicate Detection

This section delves into how researchers are currently addressing the challenge of visualizing duplicate records in clinical cancer datasets, particularly those that exclude image and video data. It explores a broad spectrum of tools, techniques, and approaches, some focused on developing new solutions, others aimed at evaluating existing systems. The reviewed studies range from interactive visual analytics platforms to scalable computational frameworks, all designed to manage the complexity of large, heterogeneous clinical datasets. To assess the effectiveness of these approaches, the review considers several key factors: the diversity of visualization techniques employed, their ability to address various dimensions of data quality, scalability to large datasets, level of user interactivity, and integration with duplicate detection algorithms. By comparing these elements, the review identifies both strengths and areas where further development is needed. A central takeaway is the pressing need for more comprehensive, user-centered visualization tools, ones that not only detect duplicates but also help users understand their broader impact on data quality. This overview sets the stage for future innovation, emphasizing the importance of creating tools that are not only technically robust but also practical and relevant for clinical use. A summary table in Table I is included to show how different studies map across these key dimensions.

A. Visualization Technique Diversity

A variety of studies have applied diverse visualization techniques, including flow diagrams, glyph-based representations, set visualizations, timelines, and interactive dashboards, to support duplicate detection and assess data quality in clinical cancer and broader biomedical datasets [1], [23], [24]. While basic formats such as tables and bar charts remain prevalent, recent surveys indicate growing interest in omics-driven and multidimensional visualizations tailored to the complexity of clinical data [3], [25]. Novel visualization techniques such as Missingness Glyphs and ensemble glyphs have been proposed to enhance pattern recognition in missing data and ensemble comparisons, respectively [26], [27]. Comparative visualization approaches for temporal and multivariate data have been developed to support patient cohort analysis and similarity assessments [20], [28].

B. Data Quality Dimension Coverage

Many studies have addressed multiple dimensions of data quality such as completeness, accuracy, consistency, and temporal stability, with particular emphasis on temporal aspects, which are especially critical in clinical and biomedical contexts [24], [29], [30]. Completeness and accuracy were the most commonly visualized data quality dimensions, particularly in studies focused on analyzing missing data patterns and detecting duplicate records [17], [21]. Temporal stability and changes over

time were captured using probabilistic change detection methods and time-oriented visualizations, allowing researchers to monitor how data quality evolves across clinical datasets [20], [30]. Some studies concentrated on specific data quality dimensions, for example, focusing on consistency in record linkage or accuracy in similarity calculations used for deduplication tasks [14], [15], [31].

C. Scalability and Performance

Scalability emerged as a key concern, with several studies showcasing visualization and deduplication techniques capable of processing millions of records, including complex datasets like clinical notes and adverse drug reaction databases [32], [33]. To improve scalability in duplicate detection pipelines, several studies leveraged parallel and distributed computing frameworks such as Apache Spark and cluster-based algorithms enabling efficient processing of large-scale clinical datasets [33], [34]. The scalability of visualization tools varied widely across studies. While some were built to handle large, heterogeneous clinical datasets, others were limited to smaller or synthetic datasets, restricting their practical application in real-world clinical environments [12], [19]. Balancing detail and overview in large datasets often involved trade-offs. To manage this, several studies used techniques like hierarchical clustering and level-of-detail visualizations, helping users navigate complex data without losing sight of important patterns or broader trends [22],[30].

D. User Interaction and Interpretability

Many visualization systems supported interactive exploration, allowing domain experts to actively engage with the data spotting anomalies, uncovering patterns in missing data, and comparing patient cohorts with greater ease and insight [15], [18], [19]. Some tools offered only limited interactivity, placing greater emphasis on automated detection and algorithmic outputs. As a result, user engagement was minimal, reducing opportunities for experts to explore data patterns or validate findings through direct interaction [1], [27].

Visual analytics approaches often incorporated user feedback and domain expertise to improve the clarity and trustworthiness of data quality assessments. By involving experts in the process, these tools became more interpretable and aligned with real-world clinical needs [4], [18]. Set-based and glyph visualizations stood out for their ability to make complex data quality issues easier to understand. Their intuitive design helped users quickly grasp patterns and anomalies, making them valuable tools for exploring duplicate records and other data inconsistencies [19], [21].

E. Integration with Deduplication Algorithms

There is a consensus that integrating visualization with deduplication and data cleansing algorithms enhances detection accuracy and facilitates iterative refinement of analytical processes [8], [12], [27], [30]. Certain frameworks integrate similarity computations directly with visualization components to provide deeper analytical insights and improve interpretability [10], [31]. Several tools offer limited interactivity, prioritizing automated detection and algorithmic outputs over active user involvement [1], [27].

Building upon this descriptive overview, the next section provides a critical analysis of the reviewed studies, highlighting their strengths, limitations, and methodological gaps. This synthesis aims to offer a clearer understanding of current practices and to outline future directions for duplicate visualization in clinical cancer datasets.

III. Evaluation of Methodological Strengths and Limitations

The existing literature on duplicate visualization in clinical cancer datasets presents a wide array of methodologies and tools, reflecting both innovative advancements and persistent challenges. Notable strengths include the development of scalable algorithms and interactive visualization techniques capable of handling complex, heterogeneous data. However, several limitations remain, particularly in the integration of data quality metrics, temporal dynamics, and user-centered design. These issues are often exacerbated by the experimental nature of many tools and the lack of comprehensive evaluations. This synthesis highlights the pressing need for more robust, sustainable, and clinically integrated visualization solutions that can effectively support duplicate detection and data quality assessment in non-image-based clinical cancer data.

A. Visualization Techniques for Duplicate Detection

A number of studies introduce innovative visualization techniques specifically designed for complex clinical datasets. These include interactive set visualizations to explore patterns of missing data and glyph-based methods to highlight data quality issues. Such approaches help users more effectively identify duplicates and anomalies within large, heterogeneous datasets [8], [19], [21]. Techniques such as Minhashing combined with Locality Sensitive Hashing have shown strong scalability, making it possible to efficiently detect duplicates within large volumes of clinical notes [27]. Despite notable progress, many visualization techniques remain insufficiently developed for the explicit detection of duplicates in clinical cancer datasets, especially those involving non-image data. The continued reliance on basic formats such as tables often limits both the interpretability and effectiveness of identifying duplicate records [3]. Moreover, the absence of standardized evaluation metrics and the limited number of comprehensive user studies hinder the ability to rigorously assess the effectiveness of these visualization methods [3], [18].

B. Integration of Data Quality Metrics

Some studies incorporate key data quality dimensions, such as completeness, accuracy, and consistency into their visualization frameworks, allowing users to more effectively identify and interpret data defects and duplicate entries [2], [8], [16]. Probabilistic frameworks for assessing temporal stability offer novel approaches to visualizing changes in data quality over time, which is particularly important for longitudinal clinical datasets [25]. However, the integration of data quality metrics into duplicate visualization tools is often limited or implicit. Many existing tools fall short of explicitly combining multiple dimensions of data quality or providing comprehensive visual analytics that address both temporal and multidimensional aspects of duplicate detection [3], [16]. The inherent complexity of clinical data, coupled with the diverse

nature of data quality issues, presents challenges that current visualization systems have yet to fully address [18].

C. Temporal and Multidimensional Visualization

Recent research has explored temporal visualization techniques such as timelines and temporal line charts to capture evolving patterns in patient data, which are relevant for understanding how duplicates emerge over time [15], [25]. Multidimensional approaches, including flow visualizations and hierarchical clustering, have also been applied to support the exploration of complex cancer registry and ensemble datasets [14], [22]. However, the explicit visualization of temporal dynamics and multidimensional characteristics of duplicates remains limited. Few systems enable direct comparisons between individual patients and cohorts or incorporate temporal stability assessments into duplicate detection workflows [15]. The high dimensionality and sparsity of clinical data further complicate effective temporal visualization, often resulting in oversimplified or fragmented representations [25].

D. Scalability and Performance of Visualization Tools

Several studies have introduced scalable algorithms and parallel processing techniques, such as Spark-based kNN classifiers and parallel deduplication methods that enable efficient handling of large clinical datasets containing millions of record [28], [29]. These approaches support real-time or near-real-time duplicate detection and visualization, which are essential for clinical applications. However, despite these advancements in scalability, many visualization tools face challenges in maintainability and lack ongoing updates, often leading to their discontinuation after initial publication [3]. The transition from prototype to routine clinical use is further hindered by resource limitations and insufficient software engineering practices, ultimately reducing the practical impact of these tools [3].

TABLE I. KEY DIMENSIONS OF STUDIES IN DUPLICATE VISUALIZATIONS

KEY DIMENSIONS	DESCRIPTIONS	STUDIES
Visualization Technique Diversity	This dimension explores how researchers visually represent duplicate records and related data quality issues in cancer datasets. Some rely on straightforward visuals like tables and bar charts, while others use more advanced methods such as glyphs, set-based diagrams, flow charts, and time-based plots. A diverse range of visualization styles allows researchers to examine duplicates from multiple perspectives, whether it's structural patterns, temporal changes, or categorical overlaps, making it easier to identify problems and gain clearer insights into the data.	[1], [15], [18], [19], [20], [26], [27]
Data Quality Dimension Coverage	This dimension focuses on how effectively visualization tools help users understand different aspects of data quality, such as completeness, accuracy, consistency, and how stable the data is over time. Rather than simply pointing out duplicates, well-designed visualizations show how those duplicates impact the overall reliability of the dataset. When multiple quality indicators are presented together, researchers gain a clearer and more comprehensive view of the data's trustworthiness.	[10],[12], [14], [15], [16], [19], [24], [25], [26]
Scalability and Performance	This dimension focuses on how well a visualization tool can manage large and complex cancer datasets, which often contain millions of records. A good system should remain responsive and easy to use, even when handling massive amounts of data. In real-world clinical environments, where data is often messy and highly varied, having a tool that performs smoothly and scales effectively is essential for making timely and accurate decisions.	[8],[14],[22],[27],[28],[29], ,[30]
User Interaction and Interpretability	This dimension focuses on how easily users can interact with visualization tools to explore and understand duplicate records. Features like zooming, filtering, comparing patient groups, and highlighting patterns allow users to engage with the data in meaningful ways. When visualizations are intuitive and responsive, they help researchers and clinicians quickly spot issues, ask relevant questions, and draw informed conclusions based on what they see.	[1], [4],[5], [6],[20], [21], [23], [24], [27],
Integration with Deduplication Algorithms	This dimension looks at how well visualization tools are integrated with computational methods for detecting duplicates, such as record linkage, similarity scoring, or machine learning. When these tools are tightly connected, users can easily visualize algorithm outputs, validate results, and adjust detection settings interactively. This integration helps bridge the gap between automated data processing and expert-driven data curation, making the entire process more transparent and user-friendly.	[1], [8], [10],[12], [27], [30], [31], [32]

E. User Interaction and Interpretability

Interactive visualization features such as dynamic filtering, selection mechanisms, and focus+context techniques play a key role in enhancing user engagement and enabling detailed exploration of duplicates and data quality issues [3], [23], [28]. Visual analytics systems that integrate data cleansing with exploratory workflows empower domain experts to iteratively uncover anomalies and patterns [23]. However, many existing tools offer only limited interactivity, often confined to basic search or hover functions, which may be inadequate for addressing the complexity of duplicate detection tasks [3]. The diversity and flexibility of visualization and user interface components also complicate usability evaluations and can hinder user adoption, particularly in multidisciplinary clinical environments. Additionally, the visualization of patient-reported outcomes and follow-up data remains an underexplored area.

F. Methodological Robustness and Evaluation

Some studies validate their duplicate detection and visualization methods through rigorous experimental evaluations, including benchmarking against large, well-curated duplicate datasets and incorporating expert assessments [1], [36], [37]. The use of probabilistic models and adaptive similarity measures has improved detection accuracy and allowed methods to better align with domain-specific characteristics of clinical data [15], [17]. However, the overall methodological rigor across studies varies considerably. Many lack comprehensive validation or rely on limited datasets, which undermines the reliability of their findings. The absence of standardized benchmarks for clinical cancer datasets and inconsistent reporting of visualization capabilities further complicate comparative evaluations and limit generalizability [3], [11]. Additionally, the inherent complexity of clinical data introduces confounding factors that challenge the robustness and consistency of duplicate detection algorithms [31].

G. Opportunities for Innovation

The literature highlights clear gaps and opportunities for advancing visualization strategies that integrate temporal stability, multidimensional data quality metrics, and user-centered interaction to improve duplicate detection and data quality assessment [3], [17], [34]. There is growing recognition of the potential to combine computational techniques with visual analytics to enhance interpretability and support clinical decision-making [4], [5], [6], [17]. Major challenges include the diversity of data types, the need for scalable solutions capable of handling large datasets, and the difficulty of embedding these tools into real-world clinical workflows. Addressing these issues will require stronger interdisciplinary collaboration yet such efforts are still relatively limited.

Across the reviewed studies, there is a broad consensus that visualization plays a vital role in helping researchers and clinicians detect and interpret duplicates within cancer and biomedical datasets. Most agree on the importance of incorporating multiple data quality indicators such as completeness and temporal stability and ensuring that tools are capable of managing large, complex datasets.

However, perspectives diverge on the extent to which visualization tools should be tightly integrated with deduplication algorithms, the diversity of visual techniques to be employed, and the level of user interaction required. These differences often reflect the distinct needs of various research domains, data types, and technical approaches, ranging from genetic sequence databases to clinical notes and electronic health records. Together, these shared understandings and differing viewpoints contribute to a more comprehensive and nuanced picture of the current landscape. They not only highlight the complexities and challenges involved in duplicate visualization within clinical cancer data, but also illuminate emerging trends, gaps in existing methodologies, and opportunities for innovation. This synthesis of perspectives serves as a foundation for identifying strategic directions for future research, refining theoretical frameworks, and enhancing the practical implementation of visualization techniques in clinical oncology settings.

IV. Theoretical and Practical Implications

This review highlights that duplicates in clinical cancer datasets are often complex and nuanced. Rather than simple repeated entries, duplicates may contain conflicting information or evolve over time, making them difficult to detect and interpret. Their inconsistent nature and varied impact on data analysis underscore the need for flexible, context-aware frameworks that can accommodate different types of duplicates and their implications for research and clinical decision [17], [36], [37]. Studies show that employing adaptive similarity measures particularly those capable of learning from the data can significantly enhance detection accuracy. These approaches move beyond traditional techniques like basic text matching, aligning with theoretical advancements in metric functional dependencies and trainable similarity functions [15], [17], [38].

The integration of temporal stability as a data quality dimension introduces a novel theoretical lens, emphasizing the dynamic nature of clinical data and the importance of probabilistic and information-geometric methods for identifying and characterizing temporal changes in duplicates and data quality [30]. Visualization theory is also evolving, recognizing the critical role of interactive and scalable visual analytics in managing heterogeneous clinical data. These tools support hypothesis generation and enable exploration of the multidimensional and temporal aspects of duplicates [19], [22], [28]. Furthermore, the literature challenges the adequacy of single benchmark evaluations, advocating for the use of multiple, diverse, and validated benchmarks to ensure the robustness and generalizability of duplicate detection models [36], [37]. Visual analytics frameworks are increasingly enriched by human-in-the-loop approaches, where visualization serves as a bridge between complex algorithmic outputs and user interpretation particularly in high-stakes settings such as clinical oncology and tumor board decision-making [3], [22].

For both clinical research and healthcare practice, the findings underscore the importance of making duplicate detection and visualization routine components of data management. These processes are essential for maintaining data quality, avoiding misleading analyses, and supporting reliable clinical decision-making, as demonstrated by tools like doppelgangR and integrated visual cleansing workflows [1], [23]. The successful deployment of scalable and parallelized algorithms [35], including those based on Minhashing, Locality Sensitive Hashing, and Spark-based kNN classifiers, illustrates the feasibility of handling large-scale clinical datasets, which is critical for real-world applications in hospital and research environments [32], [33]. Visualization tools that incorporate multidimensional, temporal, and set-based techniques enable practitioners to uncover complex patterns of duplication and missing data, thereby facilitating targeted data quality interventions and enhancing the interpretability of clinical datasets [24], [26], [39].

The adoption of interactive visualization systems that support user-driven exploration, such as temporal stability assessments and cohort comparisons has the potential to significantly enhance tumor board workflows and multidisciplinary clinical decision-making processes [3], [20]. To support this, industry and policy stakeholders should prioritize the development of comprehensive data quality monitoring infrastructures that incorporate visualization components, enabling continuous tracking and traceability of duplicates and other data quality issues within clinical data warehouses [2]. Evidence suggests that advanced visualization and duplicate detection techniques not only improve research accuracy but also contribute to patient safety, streamline healthcare operations, and support regulatory compliance. These benefits underscore the importance of investing in robust data quality tools across healthcare systems [16].

While existing visualization taxonomies often focus on general data exploration or image-based analytics, this review introduces a conceptual framework specifically tailored to the visualization of duplicate records in non-image clinical cancer datasets. Unlike traditional models that treat data quality and visualization as separate domains, our framework integrates duplicate detection algorithms, interactive visual analytics, and data quality dimensions (e.g. completeness, temporal stability) into a unified structure. This approach emphasizes human-in-

the-loop interpretation, temporal dynamics, and multidimensional data contexts areas that are underrepresented in current taxonomies. By aligning visualization strategies with specific data quality challenges posed by duplication, the framework offers a more targeted lens for evaluating and designing tools in clinical research settings.

The findings from this section further emphasize the complexity and variability of duplicate records in clinical cancer datasets. Addressing these challenges effectively requires well-structured datasets and adaptable tools capable of handling diverse duplication scenarios. To support this effort, the following section introduces key datasets that can aid researchers in developing more effective strategies for detecting and visualizing duplicates in cancer data.

V. Key Datasets for Duplicate Detection and Visualization

A range of publicly available datasets serve as valuable resources for researchers working to detect and visualize duplicates in clinical cancer data. These assets provide a foundational platform for developing more effective methods to clean, interpret, and analyze complex biomedical information:

1) The Cancer Genome Atlas (TCGA) [1] *:* The Cancer Genome Atlas (TCGA), accessible via the Genomic Data Commons (GDC), stands as one of the most comprehensive resources for cancer genomics. It integrates clinical and molecular data from 33 cancer types, encompassing thousands of patient cases and millions of data files. With its diverse range of data including whole genome sequencing, RNA-seq, methylation, and proteomic profiles, TCGA offers a powerful foundation for cancer research.

However, the scale and complexity of TCGA also present challenges for duplicate detection. The involvement of multiple contributors and data modalities increases the risk of overlapping entries, such as repeated patient identifiers or biospecimen records. While the GDC mitigates some of these risks through standardized clinical data formats and consistent bioinformatics workflows, duplicate detection remains a critical step to ensure data integrity and reliability.

TCGA also provides interactive visualization tools, such as the Cohort Builder and Mutation Frequency plots, which support the exploration of genomic alterations and clinical attributes. These tools are particularly useful for identifying anomalies or inconsistencies that may signal duplicate records, especially when combined with metadata filters and cohort analysis. Overall, TCGA is a cornerstone of cancer research, and its structured data along with integrated visualization capabilities make it an essential resource for addressing duplication challenges in large-scale biomedical datasets.

2) The Cancer Imaging Archive (TCIA) [2] *:* The Cancer Imaging Archive (TCIA) is a publicly accessible repository offering a vast collection of de-identified medical images related to cancer. These datasets are organized into collections based on disease types, such as lung cancer and imaging modalities, including MRI, CT scans, and histopathology. TCIA supports a variety of formats, with Digital Imaging and Communications in Medicine (DICOM) serving as the standard for radiological data.

Due to its aggregation of data from multiple institutions, TCIA presents both opportunities and challenges for duplicate detection. The diversity of contributors and imaging types increases the likelihood of overlapping records, such as repeated patient identifiers, imaging sessions, or processed files. Fortunately, TCIA provides rich metadata, including treatment details, patient outcomes, and genomic annotations, that can assist researchers in identifying and resolving potential duplicates.

The platform also features visualization tools and filtering options that enable users to explore imaging datasets and detect anomalies or inconsistencies. These capabilities are particularly useful for identifying redundant scans or mismatched patient data, which could compromise the accuracy of research findings. Overall, TCIA is a valuable resource for cancer imaging research, and when paired with visualization and record linkage techniques, it supports effective deduplication and enhances data quality.

3) North American Association of Central Cancer Registries (NAACCR) CiNA[3]*:* The NAACCR Cancer in North America (CiNA) Research dataset is a robust, population-based resource that consolidates cancer data from registries across the United States and Canada. It offers detailed insights into cancer incidence, survival, and prevalence, and is curated to meet high standards of completeness and data quality, making it an essential tool for cancer surveillance and public health research. While CiNA provides significant value, duplicate detection remains a challenge due to the integration of data from multiple state and provincial registries. Patients who move across regions or receive care from different institutions may have overlapping records. NAACCR addresses this through standardized coding practices and routine quality checks, but researchers must still exercise caution, particularly when working with multi-year or multi-registry datasets.

CiNA's rich metadata, including patient demographics, diagnosis timelines, and tumor characteristics, supports effective deduplication. Tools like SEER*Stat enable researchers to visualize and analyze these variables, helping to identify inconsistencies or repeated entries. This is especially important when linking CiNA with other datasets or conducting longitudinal studies, where duplicate records can distort findings. Overall, CiNA is a high-quality dataset that, when paired with appropriate visualization and data linkage techniques, facilitates accurate duplicate detection and enhances data reliability.

4) KAUH-BCMD[4]*:* The KAUH-BCMD dataset, developed by Al-Mnayyis et al. [40], comprises over 7,000

[1] https://portal.gdc.cancer.gov/
[2] https://www.cancerimagingarchive.net/
[3] https://www.naaccr.org/cina-research/
[4] https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2025.1529848/full

mammographic images from 5,000 patients at King Abdullah University Hospital in Jordan. While the primary aim of the study was to enhance breast cancer classification using deep learning techniques, the dataset's construction and preprocessing also offer valuable insights into duplicate detection within clinical imaging data.

To improve image quality, the researchers applied several preprocessing methods, including high-boost filtering, contrast-limited adaptive histogram equalization (CLAHE), normalization, and data augmentation. While these techniques are essential for optimizing model performance, they can inadvertently introduce near-duplicate images if not carefully managed, potentially affecting the accuracy of training and evaluation processes.

The KAUH-BCMD dataset also includes metadata such as patient identifiers and diagnostic labels, which can be leveraged to link records and identify potential duplicates. Although duplicate detection was not the primary focus of the original study, the dataset's size and diversity make it a promising candidate for future research in this area. With the application of appropriate visualization and record linkage techniques, KAUH-BCMD could offer valuable insights into managing duplication challenges in cancer imaging data.

5) Breast Cancer Wisconsin (UCI)[5]*:* The Breast Cancer Wisconsin (Diagnostic) dataset, hosted by the UCI Machine Learning Repository, is a widely recognized benchmark in biomedical research. It comprises 569 samples derived from digitized fine needle aspirate (FNA) images of breast tumors, with each sample described by 30 numerical features related to the shape and texture of cell nuclei, such as radius, perimeter, area, and concavity. This rich feature set makes the dataset particularly valuable for developing and evaluating machine learning models in cancer diagnostics.

In terms of duplicate detection, the dataset is exceptionally clean and well-structured, featuring unique patient identifiers and no missing values. This reliability makes it an ideal starting point for testing deduplication algorithms or building classification models without concerns about data inconsistencies. However, when used in machine learning workflows, especially during cross-validation, care must be taken to avoid data leakage or overfitting. For instance, preprocessing steps like data augmentation or synthetic sample generation can unintentionally introduce near-duplicate entries, potentially skewing model performance.

Although the dataset does not contain explicit duplicates, its structured format and detailed metadata make it suitable for evaluating deduplication techniques, particularly those focused on feature-level similarity and record linkage. Researchers can simulate duplication scenarios by introducing controlled redundancy and applying visualization methods to detect and resolve them. Overall, the Breast Cancer Wisconsin dataset serves as a valuable resource for exploring deduplication strategies in small-scale, feature-rich clinical datasets, especially within the context of diagnostic imaging and classification.

While the datasets reviewed offer strong foundations for duplicate detection and visualization in cancer research, several limitations persist. The following section outlines key gaps in current approaches and proposes future directions to advance the field.

VI. Gaps and Future Research Directions

Current visualization approaches in clinical cancer datasets tend to emphasize general data quality issues or patterns of missing data. However, they often overlook specialized techniques for detecting duplicate records, especially in non-image data. Most existing tools rely on basic visual formats like tables, which can limit users' ability to interpret and effectively identify duplicates [3], [24], [26]. To address this gap, future research could focus on developing interactive visualization tools that are specifically designed to highlight duplicates within complex clinical datasets.

These tools should be capable of integrating temporal and multidimensional data, making it easier to spot and understand duplicate entries, ultimately improving data integrity and analytical precision. Moreover, current visualization systems rarely offer a comprehensive view of multiple data quality dimensions, such as completeness, accuracy, consistency, and temporal stability within duplicate detection workflows. When these aspects are only partially or implicitly represented, users struggle to assess the full impact of duplicates on overall data quality [2], [12], [21].

There is a clear opportunity to design integrated visualization frameworks that display these metrics alongside duplicate indicators, enabling a more holistic and intuitive evaluation of clinical data. Another limitation is the lack of support for visualizing how duplicates change or cluster over time. Temporal dynamics are especially important in clinical datasets, yet they are often underrepresented in current tools [20], [30]. This opens the door for innovative solutions like interactive timelines or change detection plots, combined with multidimensional views, to track duplicates longitudinally and provide deeper insights into data evolution.

VII. Conclusions

In conclusion, this review has examined current visualization techniques for detecting duplicates in clinical cancer datasets, particularly non-image data, and identified both promising developments and critical gaps. Despite progress in scalable algorithms and interactive analytics, most tools still rely on basic visual formats and fail to fully incorporate essential data quality dimensions such as completeness, accuracy, consistency, and temporal dynamics. There is a clear need for more intuitive and clinically relevant visualization frameworks that support complex, multidimensional data and allow users to explore duplicates interactively. Effective tools should be designed with users in mind, enabling human-in-the-loop workflows and facilitating the interpretation of anomalies over time. Many existing systems lack rigorous evaluation, standardized benchmarks, and long-term usability, which limits their practical impact. Future research should focus on building integrated platforms that combine robust deduplication

[5] https://archive.ics.uci.edu/dataset/14/breast+cancer

algorithms with user-friendly, scalable visual interfaces capable of supporting longitudinal tracking, cohort comparisons, and real-time data quality monitoring. Enhancing how duplicates are visualized is essential for improving data integrity, analytical accuracy, and ultimately patient outcomes. Bridging the gap between computational detection and visual understanding will empower users to not only identify duplicates but also take meaningful action, leading to more reliable cancer research and stronger foundations for data-driven healthcare innovation.

ACKNOWLEDGMENT

The authors would like to thank the Centre for Research and Innovation Management (CRIM), UTeM, and members of the IDEAL research group for all their support. This study is funded by the Malaysian Technical University Network (MTUN) Strategic Collaboration Research Grant (MTUN/2024/UTEM-FTMK/CRG/MS0008).

REFERENCES

[1] L. Waldron, M. Riester, M. Ramos, G. Parmigiani, and M. J. Birrer, "The doppelgänger effect: Hidden duplicates in databases of transcriptome profiles," Journal of the National Cancer Institute, vol. 108, no. 11, 2016, doi: 10.1093/JNCI/DJW146.

[2] H. Spengler, I. Gatz, F. Kohlmayer, K. A. Kuhn, and F. Prasser, "Improving Data Quality in Medical Research: A Monitoring Architecture for Clinical and Translational Data Warehouses," Computer-Based Medical Systems, pp. 415–420, 2020, doi: 10.1109/CBMS49503.2020.00085.

[3] D. Boehm et al., "Data visualization support for tumor boards and clinical oncology: Scoping review," JMIR Res Protoc., vol. 5:13:e5362, 2024, doi: 10.2196/53627.

[4] C. Turkay, F. Jeanquartier, A. Holzinger, and H. Hauser, "On computationally-enhanced visual analysis of heterogeneous data and its application in biomedical informatics," 2014.

[5] E. W. Prince and T. C. Hankinson, "for Clinical Decision Support," in Pacific Symposium on Biocomputing 2025, 2025, pp. 40–53, [Online]. Available: https://doi.org/10.1093/jamia/ocaf010.

[6] J. Wu, H. Wang, C. Ni, and K. Qian, "Interactive Data Visualization Techniques for Enhancing AI Decision Transparency in Healthcare Analytics: A Comparative Analysis," Applied and Computational Engineering, vol. 146, no. 1, pp. 175–186, 2025, doi: 10.54254/2755-2721/2025.tj22322.

[7] Q. Chen, J. Zobel, and K. Verspoor, "Duplicates, redundancies and inconsistencies in the primary nucleotide databases: A descriptive study," Database, vol. 2017, no. 1, pp. 1–16, 2017, doi: 10.1093/database/baw163.

[8] W. Digan, M. Wack, V. Looten, A. Neuraz, A. Burgun, and B. Rance, Evaluating the impact of text duplications on a corpus of more than 600,000 clinical narratives in a French hospital, vol. 264. IOS Press, 2019.

[9] J. Steinkamp, J. J. Kantrowitz, and S. Airan-Javia, "Prevalence and Sources of Duplicate Information in the Electronic Medical Record," JAMA Network Open, vol. 5, no. 9, pp. 1–11, 2022, doi: 10.1001/jamanetworkopen.2022.33348.

[10] B. Connolly et al., "A Statistical Approach for Visualizing the Quality of Multi-Hospital Data," Visible Language, vol. 48, no. 3, p. 68, 2014, [Online]. Available: https://www.questia.com/library/journal/1P3-3531587361/a-statistical-approach-for-visualizing-the-quality.

[11] D. G and S. Kumaresan, "A survey on duplicate record detection in real world data," International Conference on Advanced Computing, vol. 01, pp. 1–5, 2016, doi: 10.1109/ICACCS.2016.7586397.

[12] J. M. B. Josko and J. E. Ferreira, "Vis4DD: A visualization system that supports Data Quality Visual Assessment.," in 32th Brazilian Symposium on Databases, 2017, pp. 46–51, [Online]. Available: https://sbbd.org.br/2017/wp-content/uploads/sites/3/2017/10/proceedings-satellite-events-sbbd-2017.pdf.

[13] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, "Duplicate Record Detection: A Survey," Social Science Research Network, 2006.

[14] C. Forbes, H. Greenwood, M. Carter, and J. Clark, "Automation of duplicate record detection for systematic reviews: Deduplicator," Systematic Reviews, vol. 13, no. 1, 2024, doi: 10.1186/s13643-024-02619-9.

[15] M. Bilenko and R. J. Mooney, "Adaptive duplicate detection using learnable string similarity measures," in KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data minin, 2003, pp. 39–48, doi: 10.1145/956750.956759.

[16] G. N. Norén, R. Orre, A. Bate, and I. R. Edwards, "Duplicate detection in adverse drug reaction surveillance," Data Mining and Knowledge Discovery, vol. 14, no. 3, pp. 305–328, 2007, doi: 10.1007/S10618-006-0052-8.

[17] Y. Huang and F. Chiang, "Refining Duplicate Detection for Improved Data Quality," in TDDL, MDQual and Futurity Workshops at TPDL 2017, 2017, pp. 1–10, doi: https://doi.org/10.1145/3377878.

[18] M. I. Gabr, Y. Helmy, and D. S. Elzanfaly, "Tracking The Sensitivity of The Learning Models Toward Exact and Near Duplicates," International Journal of Advanced Computer Science and Applications, vol. 13, no. 12, 2022, doi: 10.14569/ijacsa.2022.0131240.

[19] O. Bieh-Zimmert, C. Koschtial, and C. Felden, "Representing multidimensional cancer registry data," in Proceedings of the 13th International Conference on Knowledge Management and Knowledge Technologies, 2013, pp. 1–4, doi: 10.1145/2494188.2494222.

[20] J. Scheer et al., "Visualization Techniques of Time-Oriented Data for the Comparison of Single Patients With Multiple Patients or Cohorts: Scoping Review," Journal of Medical Internet Research, vol. 24, no. 10, p. e38041, 2022, doi: 10.2196/38041.

[21] R. A. Ruddle, "Using Well-Known Techniques to Visualize Characteristics of Data Quality," Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 3: VISIGRAPP, VISIGRAPP, pp. 89–100, 2023, doi: 10.5220/0011664300003417.

[22] C. Turkay, R. S. Laramee, and A. Holzinger, "On the challenges and opportunities in visualization for machine learning and knowledge extraction: A research agenda," 2017.

[23] C. Schmidt et al., "Combining Visual Cleansing and Exploration for Clinical Data," in 2019 IEEE Workshop on Visual Analytics in Healthcare (VAHC), 2019, pp. 25–32, doi: 10.1109/VAHC47919.2019.8945034.

[24] [R. A. Ruddle, M. Adnan, and M. Hall, "Using set visualization techniques to investigate and explain patterns of missing values in electronic health records," medRxiv, 2022, doi: 10.1101/2022.05.13.22275041.

[25] H. M. Shakeel, S. Iram, H. Al-Aqrabi, T. Alsboui, and R. Hill, "A comprehensive state-of-the-art survey on data visualization tools: Research developments, challenges and future domain specific visualization framework," IEEE Access, vol. 10, pp. 96581–96601, doi: 10.1109/access.2022.3205115.

[26] S. J. Fernstad and J. Johansson, "To Explore What Isn't There -- Glyph-based Visualization for Analysis of Missing Values," IEEE Transactions on Visualization and Computer Graphics, vol. 28, no. 10, pp. 3513–3529, 2021, [Online]. Available: 10.1109/TVCG.2021.3065124.

[27] L. Hao, C. G. Healey, S. A. Bass, and H.-Y. Yu, "Visualizing Static Ensembles For Effective Shape and Data Comparison.," vol. 2016, no. 1, pp. 1–10, 2016, doi: 10.2352/ISSN.2470-1173.2016.1.VDA-509.

[28] R. Guo et al., "Comparative Visual Analytics for Assessing Medical Records with Sequence Embedding," Visual Informatics, vol. 4, no. 2, pp. 72–85, 2020, [Online]. Available: https://doi.org/10.1016/j.visinf.2020.04.001.

[29] C. N. Ta and C. Weng, "Detecting systemic data quality issues in electronic health records," Studies in Health Technology and Informatics, vol. 264, pp. 383–387, 2019, doi: 10.3233/SHTI190248.

[30] C. Sáez, P. P. Rodrigues, J. Gama, M. Robles, and J. M. García-Gómez, "Probabilistic change detection and visualization methods for the assessment of temporal stability in biomedical data quality," Data Mining and Knowledge Discovery, vol. 29, no. 4, pp. 950–975, 2015, doi: 10.1007/S10618-014-0378-6.

[31] A. Waldenburger, D. Nasseh, and J. Stausberg, "Detecting duplicates at hospital admission: Comparison of deterministic and probabilistic record linkage," Studies in Health Technology and Informatics, vol. 226, pp. 135–138, 2016, doi: 10.3233/978-1-61499-664-4-135.

[32] S. Shenoy, T.-T. Kuo, R. A. Gabriel, J. McAuley, and C.-N. Hsu, "Deduplication in a massive clinical note dataset.," To be published. Accessed: Sept. 25, 2025, [Online]. Available: https://arxiv.org/pdf/1704.05617v1.pdf.

[33] C. Wang and S. Karimi, "Parallel duplicate detection in adverse drug reaction databases with spark," in Proc. 19th International Conference on Extending Database Technology (EDBT), 2016, pp. 551–562, doi: 10.5441/002/edbt.2016.52.

[34] W. J. P. dos Santos et al., "A Scalable Parallel Deduplication Algorithm," Symposium on Computer Architecture and High Performance Computing, pp. 79–86, 2007, doi: 10.1109/SBAC-PAD.2007.32.

[35] J. Schmidt, "Scalable Comparative Visualization," 2016, doi: 10.2312/2631090.

[36] Q. Chen, J. Zobel, and K. Verspoor, "Benchmarks for measurement of duplicate detection methods in nucleotide databases," Database (Oxford), 2017, doi: 10.1093/DATABASE/BAW164.

[37] Q. Chen, J. Zobel, and K. Verspoor, "Duplicates, redundancies and inconsistencies in the primary nucleotide databases: A descriptive study," Database (Oxford), 2017, doi: 10.1093/DATABASE/BAW163.

[38] W. Andrzejewski, B. Bębel, P. Boinski, and R. Wrembel, "On tuning parameters guiding similarity computations in a data deduplication pipeline for customers records," Information Systems, 2023, doi: 10.1016/j.is.2023.102323.

[39] S. Sosvilla-Rivero, "Using set visualisation to find and explain patterns of missing values: a case study with NHS hospital episode statistics data," BMJ Open, vol. 12, no. 11, p. e064887, 2022, doi: 10.1136/bmjopen-2022-064887.

[40] A. M. Al-Mnayyis et al., "(KAUH-BCMD) dataset: advancing mammographic breast cancer classification with multi-fusion preprocessing and residual depth-wise network," Frontiers in Big Data, vol. 8, 2025, doi: 10.3389/fdata.2025.1529848.