Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Advanced Techniques in Cancer Prediction: A Review of Data Mining and Machine Learning Approaches

Y. Sreenivasula Goud[1], Mukesh Kumar Tripathi[2], Ashwini Shinde[3], Deepali Bongulwar[4], Bhagyashree Tingare[5], Sanjeevkumar Angadi[6]

[1]Department of Electronics Communication and Engineering, G. Pullaiah college of Engineering and Technology, Kurnool, India.
[2]Department of Computer Science and Engineering, Vardhaman College of Engineering, Hyderabad, India.
[3]Department of Electronics and Telecommunication, Nutan Maharashtra Institute of Engineering and Technology, India.
[4]Department of CSE (Artificial Intelligence and Machine Learning), Brainware University, Kolkata, India.
[5]Department of Artificial Intelligence and Data Science, D.Y. Patil College of Engineering, Pune, India.
[6]Department of Computer Science and Engineering, Nutan College of Engineering and Research, Pune, India.

Addressing cancer disorders is a critical global health concern, especially given the challenges faced in developing countries. Insufficient prevention efforts, a lack of diagnostic tools, and limited access to trained medical professionals contribute to the complexity of diagnosis and treatment. Efforts to improve prediction and treatment outcomes through technology, mainly through clinical decision support systems (CDSS), are promising. These systems leverage data extraction and deep learning techniques to aid healthcare professionals in making informed decisions about cardiac disease. Classical machine learning algorithms such as Naïve Bayes, Decision Trees, and Artificial Neural Networks have been widely employed in predicting cardiac disease. However, there's recognition of the need for more sophisticated models that can integrate diverse sources of knowledge from different geographic regions to enhance early detection capabilities. The state of statistical models for cardiac disease prediction appears incremental, suggesting a continual need for innovation and improvement in this field. By harnessing the power of advanced technologies and collaborative efforts, we can strive towards more robust predictive models that have the potential to significantly impact the prevention and management of cancer disorders on a global scale. It's impressive that researchers developed four different models for predicting cardiac disease, utilizing a range of machine learning algorithms, including Support Vector Machine (SVM), Random Forest, Boost, and k-nearest Neighbor. Achieving high f1-scores, especially up to 88%, with SVM and Boost is a notable accomplishment and demonstrates the efficacy of these models in predicting cardiac disease.

Keywords: Cancer management, Data mining, Deep learning, Machine learning, Survival analysis.

1. Introduction

The digitization of data has indeed revolutionized cancer research by enabling the application of various data-driven practices. Data mining, which involves extracting and identifying patterns from large datasets, has become invaluable. It encompasses techniques such as association analysis, correlation analysis, classification, regression, and cluster analysis [1]. Different types of patterns can be extracted using these techniques, depending on the nature of the data and the desired outcome. The association analysis technique identifies relationships or associations between variables in the data. It is commonly used to discover frequent patterns or co-occurrences among items in large datasets, such as identifying associations between certain genes and specific types of cancer. Classification and regression methods predict outcomes or classify data into predefined categories based on input variables. In cancer research, classification algorithms can indicate the likelihood of a patient developing a particular type of cancer based on their genetic markers or other risk factors. The cluster analysis technique groups similar data points based on their characteristics or attributes. In cancer research, cluster analysis can help identify subtypes of tumours with similar molecular profiles, which may have implications for diagnosis and treatment [2-3].

Mining medical data presents both exciting opportunities and significant challenges due to the complexity and sensitivity of the data involved. In the context of cancer research, data mining techniques offer the potential to uncover intricate patterns and details within biological data, ultimately leading to better predictive models and treatment outcomes. However, it's crucial to consider the ethical implications and potential risks associated with data mining, especially patient privacy and data security. Cancer, with its high morbidity and mortality rates, remains a significant challenge to humanity [4]. However, technological advancements have enabled the application of data-driven prediction techniques to improve cancer prognosis models. Researchers can extract valuable insights into survival outcomes by analysing large datasets encompassing the previous medical history of many cancer patients. The survival outcome of cancer patients is typically measured from the time of diagnosis until the conclusion of a study or observation period. Researchers can use data mining techniques to analyse patient demographics, tumour characteristics, treatment regimens, and genetic markers to develop predictive models for cancer survival [5-6]. Over the past decades, data mining has been effectively utilized in numerous healthcare and medical applications, including cancer prognosis. By harnessing the power of data-driven approaches, researchers can enhance our understanding of cancer progression, identify prognostic factors, and tailor treatment strategies to individual patients [7].

Data mining techniques have emerged as powerful tools in cancer research, linking various clinical attributes of patients to their survival outcomes. These techniques offer high performance and can leverage a range of classification and regression algorithms to build predictive models using cancer data [8]. Researchers have employed different statistical and machine-learning techniques to identify prognostic indicators for cancer survival. To predict

patient outcomes, these techniques analyse patient-specific attributes, including demographic information and medical features. In recent years, there has been a growing trend in utilizing publicly available cancer or clinical datasets for research. Using statistical methods or machine learning algorithms, scientists and clinicians leverage these datasets to associate patient-specific characteristics with survival outcomes. Such techniques have enabled researchers to uncover valuable insights into cancer prognosis, facilitating more informed decision-making in clinical practice. By identifying prognostic factors and developing predictive models, medical practitioners can better personalize treatment plans and improve patient outcomes [9].

2. Literature Survey:

The study [10] presented a framework for predicting cancer disease using several machine learning algorithms implemented in the R programming language. Among the algorithms utilized were Support Vector Machine (SVM), k-nearest Neighbor (KNN), and Naïve Bayes (NB). The researchers employed a machine-learning repository containing 303 instances and 76 features, sourced from the Cleveland datasets available from the University of California, Irvine (UCI). The data was processed to handle missing values, resulting in a dataset of 302 samples with fourteen cancer attack characteristics. The dataset was split into training and testing sets, with 70% used for model training and 30% for testing the models' performance. The study compared the performance of the SVM, KNN, and NB classifiers in predicting cancer disease. The results indicated that the Naïve Bayes classifier achieved the highest accuracy, reaching 86.6%. This suggests that, in this study, Naïve Bayes outperformed SVM and KNN for cancer disease prediction using the given dataset and features.

The author [11] proposed a predictive method for cardiac disease using a Multi-Layer Perceptron (MLP) Neural Network, with backpropagation as the testing algorithm. The success of the developed model was evaluated based on sensitivity, precision, specificity, and accuracy metrics. For model training and validation, the researchers utilized the Cleveland dataset from the UCI Machine Learning Repository, which comprised 303 instances and 76 features. Preprocessing of the data involved removing six cases with missing values. Subsequently, only 14 out of the 76 features were identified as the most important for predicting cardiac disorders. The study reported that the MLP-NN model achieved a high precision rate of 93.39% using five hidden neurons, with a runtime of 3.86 seconds for predicting cardiac disease. This suggests that the MLP-NN model, trained on the selected features, demonstrated strong predictive performance in identifying cardiac disorders.

The author [12] introduced a machine learning approach for cardiovascular analysis, primarily focusing on logistic regression (LR). Additionally, the study investigated the efficiency of other algorithms such as Naïve Bayes (NB), Support Vector Machine (SVM), Decision Trees (DT), and k-Nearest Neighbor (KNN) by utilizing the SK-Learn library. The experimental results indicated that the LR method performed well, with a precision of 86.89%. Comparatively, other algorithms achieved the following precision rates: NB at 86%, SVM at 82%, DT at 78.69%, and KNN at 77.85%. However, it's important to note that the study did not specify the datasets for training the models or the testing processes. Overall, the author highlighted the effectiveness of logistic regression in cardiovascular analysis, demonstrating high precision compared to other machine learning algorithms. However, further details

regarding the dataset characteristics and experimental methodology would provide additional context for interpreting the results.

The author [13] conducted a similar study focusing on the predictive value of cardiovascular disease using Artificial Neural Network (ANN) and Support Vector Machine (SVM) classification algorithms. The dataset was collected from three designated clinics affiliated with AJA Medical Sciences University in Iran, encompassing 1324 occurrences and 25 attributes. The data comprised historical reports of patients hospitalized for coronary cancer disease between March 2016 and March 2017. The variables used in the study were derived from the UCI Machine Learning Repository Cleveland Data Guideline. Data collection involved various processes: analysis, cleansing, normalization, and reduction. The dataset was split into 70% for training and 30% for testing the algorithms. Experimental results indicated that the SVM algorithm exhibited higher reliability, effectiveness, strength, and sensitivity than the ANN model. This suggests that SVM may be a more suitable approach for predicting cardiovascular disease in this dataset and context.

The author [14] analyzes cardiovascular disease risk prediction using automatic machine learning compared to a graduate student implementing various machine learning algorithms. The study focused on critical parameters such as the time required for constructing machine learning models and the accuracy of predictions on unseen datasets. The Auto-SKlearn model, inspired by the widely-used Scikit-Learn toolbox, was employed. This model incorporates several classifications and preprocessing techniques available in the Scikit-Learn toolbox, including Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Boosting, Neural Networks (NN), and k-Nearest Neighbors. The results indicated that the AutoML model was significantly faster than the manual implementation by the graduate student. It took only 30 minutes to create classifiers using AutoML, whereas the student spent 431 hours for the UCI dataset and 360 hours for the larger dataset. However, despite its efficiency, the AutoML model may have limitations in its effectiveness and performance compared to manually-tuned models, especially in more complex or nuanced scenarios.

A machine-based simulation platform for diagnosing cardiomyopathy, explicitly focusing on handling unbalanced datasets using various sampling techniques [15]. The study employed random sampling, Synthetic Minority Over-sampling Technique (SMOTE), and Adaptive Synthetic Sampling Method (ADASYN). The Framingham dataset from the Kaggle database was utilized to train and evaluate algorithms. This dataset consists of 4239 instances with 15 attributes, and the objective was to predict the probability of a patient experiencing a coronary cancer attack within ten years. Several machine learning techniques were employed, including Logistic Regression (LR), k-nearest Neighbors (KNN), AdaBoost, Decision Trees (DT), Naïve Bayes (NB), and Random Forest (RF). These classification algorithms were evaluated based on metrics such as accuracy, precision, and recall, with their performance influenced by the sampling techniques used. The experimental results revealed that the Support Vector Machine (SVM) with a Random Over-Sampling Technique achieved the highest precision of 99% in predicting cancer disease. However, with the SMOTE technique, Random Forest (RF) performed best with an accuracy of 91.3%. In comparison, both Decision Trees (DT) and Random Forests (RF) achieved a precision of 90.3% with the ADASYN technique.

The author [16] conducted a comparative analysis between Decision Trees (DT) and Support

Vector Machine (SVM) classification algorithms for implementing a machine learning approach aimed at cancer disease prevention. The study utilized Python for implementation. The dataset included features related to cardiac disease, such as age, chest pain, blood pressure, and cholesterol levels. Data preprocessing involved eliminating inconsistencies and missing values using the PANDAS formula. Additionally, data visualization was performed using the Matplotlib library for better understanding. Experimental results indicated that the DT classification algorithm achieved a significantly higher accuracy than the SVM algorithm. Specifically, the DT classifier achieved a perfect classification accuracy of 100%, while the SVM classifier achieved an accuracy of only 55%. The study argued that the classification success depends on the characteristics of the cancer disease dataset. While the DT classifier demonstrated exceptional performance in this dataset, achieving 100% accuracy, it cannot be generalized as the best predictor for cancer disease prediction across all datasets. Different datasets may exhibit varying characteristics and complexities, influencing the performance of various classification algorithms.

3. Background

Cancer research encompasses many topics, but prognosis prediction is crucial in understanding and managing the disease. Prognosis prediction in cancer typically involves three main areas: cancer diagnosis, cancer survival prediction, and cancer recurrence prediction. Each area focuses on different aspects of the disease and predicts outcomes at various time points.

3.1 Cancer Diagnosis:

Cancer diagnosis focuses on identifying whether a patient has cancer or not, as well as determining the type and stage of the cancer. Diagnostic tools such as imaging techniques, biopsies, and molecular tests are used to detect and classify cancerous tumours. Cancer diagnosis aims to initiate timely treatment and intervention to improve patient outcomes. Researchers have explored novel diagnostic techniques to address this challenge, including those based on data mining and machine learning in clinical research. These techniques leverage various parameters, gene biomarkers, imaging data such as CT scans, and other relevant information to improve cancer detection and classification. Machine learning techniques offer the potential to analyse large and complex datasets, identify patterns, and make predictions that may not be apparent to human observers. By training models on diverse datasets containing information from cancer patients, researchers can develop algorithms capable of accurately detecting and classifying cancer in different body parts. These machine-learning models can complement traditional diagnostic methods and potentially enhance the efficiency and accuracy of cancer diagnosis. They may help identify subtle patterns or biomarkers indicative of early-stage cancer, enabling earlier detection and intervention. Integrating data mining and machine learning techniques into clinical research holds promise for advancing cancer diagnosis and reducing mortality rates by improving early detection and treatment outcomes. Continued research in this field is essential to refine further and validate these innovative approaches in real-world clinical settings [17-18].

3.2 Cancer Survival Prediction:

Cancer survival prediction involves estimating the likelihood of a patient surviving the disease

over a specified period, typically after receiving a cancer diagnosis and undergoing treatment. This prediction may consider various factors such as tumour characteristics, treatment options, patient demographics, and genetic markers. Survival prediction models help clinicians and patients make informed decisions about treatment strategies and care plans. The evolution of survival prediction in cancer has indeed undergone significant advancements over the past few decades, as depicted in Fig. 1. In the early days, clinicians relied primarily on their experience and knowledge to forecast the survival of cancer patients, which was considered an ideal approach at the time. However, with the emergence of advanced technologies and the integration of information technology (IT) developments in the medical domain, there has been a significant shift towards more sophisticated methods of survival prediction, marking a clear progress in the field. One notable development is the utilization of nomograms, which are widely used tools in oncology [19-20]. Nomograms involve the use of statistical models to compute the probability of survival or recurrence in cancer patients based on various parameters such as tumour grade, patient age, and other clinical characteristics. These models provide a quantitative and personalized approach to survival prediction, allowing clinicians to make more informed decisions about treatment strategies and patient care. Overall, the adoption of nomograms and statistical models has revolutionized survival prediction in cancer by providing more accurate and individualized estimates of patient outcomes. This evolution reflects the ongoing progress in oncology research and clinical practice, ultimately leading to improved patient care and management.

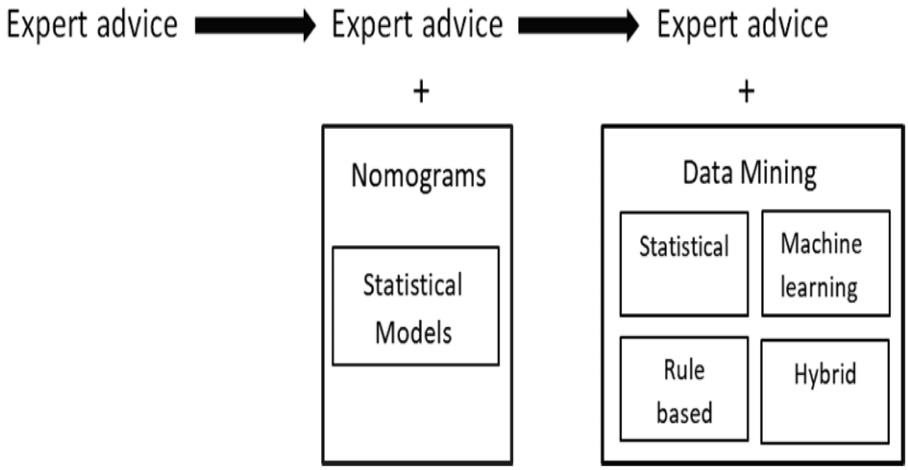

Fig 1: Cancer survival prediction: from past to present

3.3 Cancer Recurrence Prediction:

Cancer recurrence prediction, a crucial aspect of cancer management, focuses on identifying the likelihood of cancer returning or progressing after initial treatment and remission. This prediction is vital for monitoring patients' post-treatment and determining the need for additional interventions or surveillance. Factors such as tumour biology, treatment response, and patient-specific risk factors are considered in recurrence prediction models. While these areas of prognosis prediction in cancer share similarities, they differ in terms of the time frame involved and the specific outcomes being predicted. Cancer diagnosis focuses on the initial

detection and classification of cancer, while cancer survival prediction estimates the likelihood of long-term survival following diagnosis and treatment. On the other hand, cancer recurrence prediction assesses the risk of cancer returning or progressing after initial treatment. By addressing each of these aspects, researchers and clinicians can better understand and manage cancer, offering hope for improved patient outcomes and quality of life.

4. Methods:

Using the PRISMA methodology for our literature review is a robust approach for systematically gathering and synthesizing relevant research findings. By searching multiple repositories like PubMed, IEEE Xplore, Springer, and Science Direct, you ensure comprehensive coverage of studies published over a significant period from 2006 to 2024. Including studies published up to 2024 allows you to capture past trends, current practices, and emerging techniques used by researchers in predicting cancer survival. This approach enables you to provide a comprehensive overview of the methods and dynamics involved in this field. It's also noteworthy that our review encompasses a broad spectrum of cancer types rather than focusing on a specific type. This wide scope allows for a more generalized understanding of the predictive methods and trends across different types of cancer, which can be valuable for researchers and practitioners working in oncology.

Our search strategy is meticulously designed to capture the most relevant literature on cancer survival prediction. By employing a combination of disease-related terms (such as "Cancer"), task-related terms (such as "Survival" and "Prediction" or "Prognosis"), and techniques-related terms (such as "Data mining," "Machine learning," "Classification," "Rule mining," and "Sequence mining"), we effectively narrow down the search to articles explicitly addressing the prediction of cancer survival using various data mining and machine learning techniques. After retrieving a substantial number of publications (378) from multiple databases, the crucial step of removing duplicate articles is undertaken to ensure that each unique study is counted only once. This meticulous process helps eliminate redundancy and ensures that the final dataset of articles for analysis is clean and accurate. After removing duplicates, we have a refined dataset of 569 articles for further analysis. This comprehensive collection of studies on cancer survival prediction using data mining and machine learning techniques serves as a rich source of information for our literature review and analysis. Overall, our search strategy and subsequent data processing steps are thorough and systematic, laying a solid foundation for conducting a comprehensive literature review.

The literature review is structured to comprehensively address the different aspects involved in survival prediction using data mining techniques, as illustrated in Fig. 2. By breaking down the process into four main aspects and further delineating seven sub-processes within it, we provide a systematic framework for analysing and comparing the research articles. The manual analysis of each article to identify key elements such as the dataset used, cancer type studied, feature selection methods, classification techniques, and validation measures is a meticulous approach. This level of detail allows for a thorough examination of the methodologies employed in each study and facilitates comparisons between different research findings. By systematically addressing each sub-process, our literature review can provide insights into the current trends, techniques, and challenges in survival prediction research. Additionally, it

enables us to identify gaps in the existing literature and areas where further research is needed. Overall, our approach to the literature review is comprehensive and well-organized, providing a valuable contribution to the understanding of survival prediction using data mining in the context of cancer research.

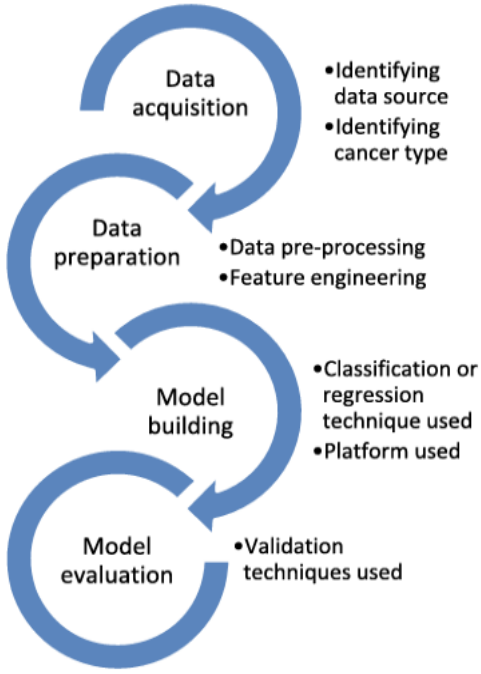

Fig 2: Data mining steps for survival prediction.

5. Open issues and possible future aspects:

The increasing complexity of healthcare, coupled with the rising burden of cancer worldwide, underscores the urgent need for effective strategies to address the challenges faced by patients and clinicians. According to the World Health Organization (WHO), the number of deaths due to cancer has been steadily increasing, reaching around 10 million in 2024, highlighting the significant impact of this disease on global health [22]. While researchers have made considerable efforts to develop survival estimation models for cancer, there remain several drawbacks and issues that need to be addressed to mitigate the burden of cancer. Some of these challenges include:

- Accuracy and Reliability: Many existing models may lack accuracy and reliability, leading to inaccurate predictions of survival outcomes for cancer patients. Improving the robustness and precision of these models is essential for better patient management and decision-making.

- Generalizability: Some models may not be generalizable across different populations or cancer types, limiting their utility in diverse clinical settings. Developing models that can be applied across various patient demographics and cancer types is crucial for widespread adoption and effectiveness.

- Interpretability: The interpretability of survival estimation models is vital for clinicians to understand and trust the predictions provided. Enhancing the transparency and interpretability of these models can facilitate their integration into clinical practice.

- Integration into Clinical Workflow: Incorporating survival estimation models into routine clinical practice can be challenging. Ensuring seamless integration with existing healthcare systems and workflows is essential to maximize their impact on patient care.

- Ethical and Regulatory Considerations: Addressing ethical and regulatory considerations, such as patient privacy, data security, and regulatory compliance, is critical to safeguarding patient rights and ensuring responsible use of predictive models in healthcare.

- To tackle these issues and advance the field of cancer survival prediction, future research agendas should focus on:

- Improving Model Performance: Continuously refining and enhancing the performance of survival estimation models through the integration of advanced machine learning techniques, larger and more diverse datasets, and rigorous validation methods.

- Enhancing Model Interpretability: Developing interpretable models that provide transparent explanations for their predictions, allowing clinicians to understand the underlying factors influencing survival outcomes.

- Addressing Healthcare Disparities: Investigating and addressing healthcare disparities to ensure equitable access to cancer care and survival prediction tools among diverse populations.

- Longitudinal Data Analysis: Incorporating longitudinal data analysis to capture changes in patient health status over time and provide dynamic predictions of survival outcomes.

- Collaboration and Data Sharing: Promoting collaboration among researchers, clinicians, and healthcare organizations to facilitate data sharing, standardization, and replication of findings across different settings.

6. Dataset availability:

The challenges associated with limited datasets in survival analysis for cancer and the ethical issues surrounding data sharing in healthcare are indeed significant barriers to progress in the field [23]. Here are some potential strategies and recommendations to address these challenges:

- Ethical Data Sharing: Establishing protocols and frameworks for ethical data sharing in healthcare can enable medical practitioners and researchers to collaborate effectively while safeguarding patient privacy and confidentiality. By anonymizing patient data and removing personal details, researchers can mitigate ethical concerns and facilitate the sharing of clinical datasets for research purposes.

- Collaboration between Clinicians and AI Specialists: Encouraging collaboration between clinicians and AI specialists can enhance the development of reliable and robust machine learning models for survival analysis in cancer. By leveraging the expertise of both domains,

interdisciplinary teams can design and implement innovative solutions that improve decision-making in healthcare.

- Promoting Reproducibility: Promoting reproducibility in research by making datasets and methodologies openly available can enhance transparency and reliability in survival analysis studies. Researchers can document their data processing and analysis pipelines, making it easier for others to replicate their findings and validate their results.

- Open Data Initiatives: Supporting open data initiatives such as OHDSI (Observational Health Data Sciences and Informatics) can foster collaboration and innovation in healthcare decision-making. By providing access to standardized and curated healthcare datasets, these initiatives enable researchers to develop and evaluate machine learning models on diverse patient populations and clinical settings.

- Policy and Regulation: Implementing policies and regulations that encourage responsible data sharing and promote transparency in healthcare research can help overcome barriers to collaboration and reproducibility. By establishing clear guidelines for data governance and sharing, policymakers can create an environment conducive to ethical and impactful research in healthcare.

7. Use of Bio-inspired computing approaches:

Incorporating meta-heuristic techniques into cancer survival prediction studies presents an exciting opportunity to enhance the performance and robustness of predictive models. Meta-heuristic algorithms, such as Particle Swarm Optimization (PSO), Cuckoo Search, Flower Pollination Algorithm, and others, offer powerful optimization capabilities that effectively address complex feature selection and model optimization challenges in survival analysis [24]. By leveraging meta-heuristic algorithms, researchers can explore high-dimensional feature spaces, identify relevant prognostic factors, and optimize machine learning models for improved predictive accuracy. These techniques can complement traditional machine learning approaches and enhance performance by guiding the search process toward optimal solutions in large and heterogeneous datasets. As a future research agenda, exploring diverse meta-heuristic algorithms, such as Cuckoo Search, Flower Pollination Algorithm, Genetic Algorithms, and Simulated Annealing, holds significant potential for advancing cancer survival prediction studies. By systematically evaluating and comparing these algorithms in the context of survival analysis, researchers can identify the most effective approaches for different types of cancer and clinical scenarios. Furthermore, integrating meta-heuristic algorithms with state-of-the-art machine learning techniques, such as deep learning, ensemble methods, and support vector machines, can further enhance the predictive capabilities of survival prediction models. This interdisciplinary approach offers exciting opportunities for innovation and discovery in cancer research, ultimately leading to improved patient outcomes and personalized treatment strategies.

8. Hybrid approaches with machine learning:

The integration of hybrid approaches and emerging technologies such as the Internet of Things

(IoT) into survival prediction models holds significant promise for improving patient outcomes and advancing healthcare decision-making [25-26]. Hybrid methods combine multiple approaches or techniques to leverage their strengths and mitigate their weaknesses. In the context of survival prediction, hybrid methods can integrate various data preprocessing techniques, feature selection algorithms, and machine learning models to improve predictive performance and address quality issues in datasets. By combining complementary methodologies, hybrid methods can enhance the robustness and accuracy of survival prediction models. IoT technology enables real-time monitoring of patients' health status using connected devices and sensors. Wearable devices, such as smartwatches and fitness trackers, can collect physiological data such as heart rate, blood pressure, and activity levels, providing valuable insights into patients' health conditions. By integrating IoT data with machine learning algorithms, researchers can develop predictive models that incorporate real-time health monitoring data to predict patient outcomes more accurately. While IoT and machine learning offer significant potential for improving survival prediction, it is crucial to address privacy and security concerns associated with the collection and storage of sensitive patient data. Robust encryption techniques, secure data transmission protocols, and adherence to privacy regulations can help safeguard patients' privacy and ensure the ethical use of their health data in predictive modeling. Reinforcement learning and case-based reasoning are advanced AI techniques that can enhance prediction models' overall accuracy by enabling adaptive learning and decision-making. These approaches allow models to learn from experience and dynamically adjust their predictions based on new information or feedback, leading to more personalized and effective patient care. By incorporating hybrid methods, IoT technology, and advanced AI techniques into survival prediction models, researchers can develop more accurate, reliable, and patient-centered predictive models that improve clinical decision-making and ultimately enhance patient outcomes. However, it is essential to address privacy concerns, ensure data security, and continue refining these approaches through rigorous validation and testing in clinical settings.

9. Conclusion:

In conclusion, a comprehensive review of cancer survival prediction using data mining techniques provides valuable insights into the field's current state and highlights important areas for future research. The SEER database emerged as the most widely used data source for cancer survival prediction studies. However, there is a need to analyse other cancer types besides breast and lung cancer to ensure comprehensive coverage and applicability of predictive models across different cancer types. While feature selection techniques are commonly used to identify critical features in datasets, there needs to be more focus on imputation techniques to handle missing data. Addressing missing data issues is crucial for improving the reliability and accuracy of predictive models. Clinicians and researchers are increasingly adopting machine learning techniques, including neural networks, ensemble approaches, and deep learning techniques, for classification tasks in cancer survival prediction. These techniques offer advanced capabilities for modelling complex relationships within the data and improving predictive performance. Various validation techniques, such as k-fold cross-validation and the holdout method, are widely used to validate predictive models. Ensuring robust validation of models is essential for assessing their generalizability and

reliability in real-world clinical settings. Open-source tools like Weka and R packages are commonly utilized for analysis in cancer survival prediction studies. Leveraging these tools enables researchers to access various algorithms and functionalities for model development and evaluation. The review identified ten challenges in the existing literature, including data quality, model interpretability, and scalability issues. Addressing these challenges is crucial for advancing the field and improving the effectiveness of predictive models in clinical practice. Future research directions should address the identified challenges and explore novel approaches, such as hybrid methods, meta-heuristic techniques, and IoT integration, to enhance predictive performance and reliability. Additionally, there is a need for greater collaboration between researchers and clinicians to ensure the practical relevance and usability of predictive models in clinical decision-making. Our comprehensive review serves as a roadmap for new researchers seeking to understand the current state of the art and existing researchers looking to identify critical issues and prospects for improving cancer survival prediction using data mining techniques. By addressing the identified challenges and embracing innovative approaches, researchers can make significant strides toward improving patient outcomes and advancing personalized cancer care.

References

1. Kirola, M., Memoria, M., Dumka, A., & Joshi, K. (2022). A comprehensive review study on: optimized data mining, machine learning and deep learning techniques for breast cancer prediction in big data context. Biomedical and Pharmacology Journal, 15(1), 13-25.
2. Islam, R., Sultana, A., & Islam, M. R. (2024). A comprehensive review for chronic disease prediction using machine learning algorithms. Journal of Electrical Systems and Information Technology, 11(1), 27.
3. MAhsan, M. M., Luna, S. A., & Siddique, Z. (2022, March). Machine-learning-based disease diagnosis: A comprehensive review. In Healthcare (Vol. 10, No. 3, p. 541). MDPI.
4. Lin, Y., Liang, R., Qiu, Y., Lv, Y., Zhang, J., Qin, G., ... & Mao, Y. (2019). Expression and gene regulation network of RBM8A in hepatocellular carcinoma based on data mining. Aging (Albany NY), 11(2), 423.
5. Shivendra, Chiranjeevi, K., & Tripathi, M. K. (2022). Detection of fruits image applying decision tree classifier techniques. In Computational Intelligence and Data Analytics: Proceedings of ICCIDA 2022 (pp. 127-139). Singapore: Springer Nature Singapore.
6. Pal, P., & Taqi, S. A. A. (2020). Advancements in Data Mining and Machine Learning Techniques for Predicting Human Diseases: A Comprehensive Review. International Journal of Research in Informative Science Application & Techniques (IJRISAT), 4(11), 19-35.
7. Behera, M., Fowler, E. E., Owonikoko, T. K., Land, W. H., Mayfield, W., Chen, Z., ... & Heine, J. J. (2011). Statistical learning methods as a preprocessing step for survival analysis: evaluation of concept using lung cancer data. Biomedical engineering online, 10, 1-15..
8. Tripathi, M. K., Maktedar, D., Vasundhara, D. N., Moorthy, C. V. K. N. S. N., & Patil, P. (2023). Residual life assessment (RLA) analysis of apple disease based on multimodal deep learning model. International Journal of Intelligent Systems and Applications in Engineering, 11(3), 1042-1050.
9. Ganggayah, M. D., Taib, N. A., Har, Y. C., Lio, P., & Dhillon, S. K. (2019). Predicting factors for survival of breast cancer patients using machine learning techniques. BMC medical informatics and decision making, 19, 1-17.
10. Zhao, B., Gabriel, R. A., Vaida, F., Lopez, N. E., Eisenstein, S., & Clary, B. M. (2020).

Predicting overall survival in patients with metastatic rectal cancer: a machine learning approach. Journal of Gastrointestinal Surgery, 24(5), 1165-1172.
11. Wang, Y., Wang, D., Ye, X., Wang, Y., Yin, Y., & Jin, Y. (2019). A tree ensemble-based two-stage model for advanced-stage colorectal cancer survival prediction. Information Sciences, 474, 106-124.
12. Tripathi, M. K., & Shivendra. (2024). Neutrosophic approach based intelligent system for automatic mango detection. Multimedia Tools and Applications, 83(14), 41761-41783.
13. Pati, D. P., & Panda, S. (2020). A Comprehensive Review on Cancer Detection and Prediction Using Computational Methods. In Computational Intelligence in Data Mining: Proceedings of the International Conference on ICCIDM 2018 (pp. 629-640). Springer Singapore.
14. Murthy, N. S., & Bethala, C. (2023). Review paper on research direction towards cancer prediction and prognosis using machine learning and deep learning models. Journal of Ambient Intelligence and Humanized Computing, 14(5), 5595-5613.
15. Tripathi, M. K., Reddy, P. K., & Neelakantappa, M. (2023). Identification of mango variety using near infrared spectroscopy. Indonesian Journal of Electrical Engineering and Computer Science, 31(3), 1776-1783.
16. Sharma, A., & Rani, R. (2021). A systematic review of applications of machine learning in cancer prediction and diagnosis. Archives of Computational Methods in Engineering, 28(7), 4875-4896.
17. Lee, M. (2023). Deep learning techniques with genomic data in cancer prognosis: a comprehensive review of the 2021–2023 literature. Biology, 12(7), 893.
18. Sharma, T., & Shah, M. (2021). A comprehensive review of machine learning techniques on diabetes detection. Visual Computing for Industry, Biomedicine, and Art, 4(1), 30.
19. Yaqoob, A., Aziz, R. M., Verma, N. K., Lalwani, P., Makrariya, A., & Kumar, P. (2023). A review on nature-inspired algorithms for cancer disease prediction and classification. Mathematics, 11(5), 1081.
20. Yaqoob, A., Aziz, R. M., Verma, N. K., Lalwani, P., Makrariya, A., & Kumar, P. (2023). A review on nature-inspired algorithms for cancer disease prediction and classification. Mathematics, 11(5), 1081.
21. Rajeashwari, S., & Arunesh, K. (2023, September). Chronic disease diagnosis and classification using data mining approaches-a comprehensive review. In AIP Conference Proceedings (Vol. 2831, No. 1). AIP Publishing.
22. Singh, P., Tripathi, M. K., Patil, M. B., Shivendra, & Neelakantappa, M. (2024). Multimodal emotion recognition model via hybrid model with improved feature level fusion on facial and EEG feature set. Multimedia Tools and Applications, 1-36.
23. Khan, F. A., Zeb, K., Al-Rakhami, M., Derhab, A., & Bukhari, S. A. C. (2021). Detection and prediction of diabetes using data mining: a comprehensive review. IEEE Access, 9, 43711-43735.
24. Tripathi, M. K., Neelakantapp, M., Kaulage, A., Nabilal, K. V., Patil, S. N., & Bamane, K. D. (2023). Breast cancer image analysis and classification framework by applying machine learning techniques. International Journal of Intelligent Systems and Applications in Engineering, 11(3), 930-941.
25. Ghorbani, R., & Ghousi, R. (2019). Predictive data mining approaches in medical diagnosis: A review of some disease's prediction. International Journal of Data and Network Science, 3(2), 47-70.
26. Zhang, S., Bamakan, S. M. H., Qu, Q., & Li, S. (2018). Learning for personalized medicine: a comprehensive review from a deep learning perspective. IEEE reviews in biomedical engineering, 12, 194-208.