IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 37, No. 5, October, 2023, pp. 1273-1280
Journal homepage: http://iieta.org/journals/ria

A Systematic Review of Machine Learning Prediction Models for Colorectal Cancer Patient Survival Using Clinical Data and Gene Expression Profiles

Ernest E. Onuiri*, Oyebola Akande, Olamide B. Kalesanwo, Taiwo Adigun, Kehinde Rosanwo, Kelechi C. Umeaka

Department of Computer Science, Babcock University, Ilishan-Remo 121103, Ogun State, Nigeria

Corresponding Author Email: onuirie@babcock.edu.ng

https://doi.org/10.18280/ria.370520

Received: 29 August 2023
Revised: 25 September 2023
Accepted: 2 October 2023
Available online: 31 October 2023

Keywords:
clinical data, colorectal cancer, gene expression, machine learning, patient survival, predictive modelling

ABSTRACT

Colorectal cancer persistently ranks among the top causes of cancer-related mortality globally. The development of superior predictive methodologies is imperative for augmenting survival outcomes. This systematic review, conducted in accordance with PRISMA-P guidelines, scrutinizes studies carried out between 2013 and 2023 that apply machine learning models to prognosticate survival in colorectal cancer patients, particularly those models incorporating clinical data and gene expression profiles. Criteria for inclusion comprised studies employing machine learning techniques, with specific emphasis on those integrating clinical data and gene expression profiles for predictive purposes. Studies devoid of explicit methodological delineation or not written in English were excluded. Decision trees, neural networks, and support vector machines emerged as the most frequently scrutinized models in the review. While some models manifested high accuracy, others underscored areas requiring refinement. Predominant data sources included patient clinical records, gene expression datasets, and molecular profiling. The results underscore the potential of machine learning in bolstering predictive precision, thereby implicating a trajectory for future research targeting the optimization of patient prognosis and treatment outcomes in colorectal cancer.

1. INTRODUCTION

Globally, cancer remains a formidable health challenge, with projections indicating that approximately 20 million individuals received a cancer diagnosis, and tragically, 10 million succumbed to the disease. The next two decades anticipate a 60% surge in cancer cases, intensifying the pressure on healthcare infrastructures, especially in low- to middle-income countries [1]. To combat this escalating crisis, the implementation of evidence-based strategies encompassing cancer prevention, early detection, and treatment is paramount. Several modifiable risk factors, such as tobacco consumption, limited intake of fruits and vegetables, excessive alcohol use, and physical inactivity, have been identified as significant contributors to cancer's prevalence.

Colorectal cancer (CRC) stands out as a major global health concern, accounting for roughly 10% of all cancer diagnoses [1]. In the United States alone, CRC is the second leading cause of cancer-related fatalities, with projections for 2021 estimating 52,980 deaths and 149,500 new diagnoses [2-4]. The gravity of CRC underscores the urgency for enhanced screening and early detection methodologies [5]. Fortunately, the past decades have witnessed a rise in survival rates, attributed in part to technological advancements, notably in the realms of artificial intelligence (AI) and machine learning (ML). These ML algorithms, capable of processing vast medical datasets, have revolutionized diagnostic accuracy, personalized treatment strategies, and patient monitoring [1, 6].

However, the application of ML in predicting survival outcomes for various cancers, including CRC, is still in its nascent stages, necessitating a systematic review to assess the quality and robustness of existing prediction models [1, 7, 8]. The potential of ML in forecasting CRC patient prognosis using gene expression profiles is evident. Furthermore, integrating clinical and radiomic attributes can further enhance the prediction accuracy for CRC patient survival [9, 10].

Diving deeper into CRC's etiology, a myriad of risk factors emerges. Environmental determinants like obesity, sedentary lifestyles, smoking, alcohol consumption, and dietary choices play a pivotal role. Concurrently, genetic factors, such as familial CRC history and specific inherited genetic mutations, amplify the risk [2, 11]. The insidious nature of CRC, often remaining asymptomatic in its initial stages, makes early detection through screening indispensable. Symptoms like rectal bleeding and abdominal discomfort typically manifest in advanced disease stages, emphasizing the criticality of proactive screening, especially for high-risk groups and individuals over 50 [11-13].

The integration of AI in oncology promises enhanced diagnostic precision and expedited clinical decision-making, culminating in improved patient outcomes. AI's potential to bridge health disparities, especially in resource-constrained settings, is noteworthy. Recognizing this potential, the National Cancer Institute champions AI endeavors, investing in research, infrastructure, and workforce development [6, 14-16]. The quest for accurate cancer patient survival predictions is pivotal for informed prognostic discussions and treatment planning. Current prediction models for CRC exhibit limitations, given the diverse survival outcomes stemming from a spectrum of molecular characteristics. This diversity

underscores the pressing need for sophisticated prediction models that encapsulate these nuances, offering individualized prognoses [5, 17, 18].

The crux of this systematic review is to provide a comprehensive overview of contemporary research on ML techniques' efficacy in predicting CRC patient survival using clinical data and gene expression profiles [6, 19-21]. This synthesis aims to inform future research trajectories and potentially refine clinical practices, enhancing CRC patient prognosis and outcomes.

1.1 Rationale

This study's genesis lies in the aspiration to systematically review the extant literature on ML models' utility in predicting CRC patient survival using clinical data and gene expression profiles. The overarching goal is to discern the most potent models and evaluate their performance metrics, including accuracy, sensitivity, and specificity.

1.2 Objectives

This study's primary objective is a meticulous review of the prevailing literature on ML models' application in predicting CRC patient survival, leveraging clinical data and gene expression profiles. The endeavor also seeks to identify the most efficacious ML models for this purpose and assess their performance metrics. Employing the PICOS framework, the study addresses pivotal research questions:

1 Population: What attributes define the patient cohort in studies that employ ML models for CRC survival prediction?
2 Intervention: Which specific ML models are predominant in these studies, and what types of clinical and gene expression data serve as their foundation?
3 Comparison: How do various ML models fare against each other in terms of their predictive capabilities using the aforementioned data?
4 Outcome: Among the myriad of ML models, which ones emerge as the most effective in predicting CRC survival, and what are their respective performance metrics?
5 Study Design: Assessing the quality and potential biases of studies that have harnessed ML models for CRC survival prediction is crucial. How do these studies measure up in terms of rigor, and what biases might influence their outcomes?

2. METHODS

In alignment with the PRISMA guidelines [22], an exhaustive literature search was executed on Scopus and PubMed databases. The search encompassed articles published from January 2013 to April 28, 2023, in English and subjected to peer review. The search strategy incorporated keywords such as "machine learning," "artificial intelligence," "predictive modelling," "colorectal cancer," "patient survival," "clinical data," and "gene expression."

2.1 Scope of the review

This systematic review encompasses studies that employed any machine learning technique to predict patient survival in colorectal cancer using clinical data and gene expression profiles.

2.2 Eligibility criteria

The PICOS framework informed the eligibility criteria:

Participants: Studies involving colorectal cancer patients.

Interventions: Research employing artificial intelligence/machine learning models for patient survival prediction using clinical and gene expression data.

Comparisons: Studies contrasting machine learning model performance with conventional survival prediction methods.

Outcomes: Research reporting prognostic accuracy metrics of machine learning models, such as sensitivity, specificity, positive predictive value, and negative predictive value.

Study Design: Observational studies, clinical trials, or simulation studies were considered.

2.3 Inclusion criteria

a. Research focusing on colorectal cancer patient survival prediction using machine learning models, incorporating clinical data and gene expression profiles.
b. Studies with colorectal cancer patients as the primary cohort.
c. Research reporting accuracy metrics of the prediction models.
d. English language publications.
e. Articles published from January 2013 to May 2023.

2.4 Exclusion criteria

a. Research not employing machine learning models with clinical data and gene expression profiles for colorectal cancer patient survival prediction.
b. Studies with primary cohorts of patients with other cancer types.
c. Research not reporting prediction model accuracy metrics.
d. Non-English publications.
e. Inaccessible full-text articles.
f. Conference abstracts, letters, editorials, case reports, reviews, and meta-analyses.

2.5 Information sources

A meticulous search was orchestrated on Scopus and PubMed databases to identify pertinent articles. The search strategy amalgamated medical subject headings (MeSH) terms and keywords pertinent to machine learning, colorectal cancer, clinical data, gene expression data, and survival prediction. This strategy was tailored to each database's unique specifications to ensure a comprehensive search [23, 24].

2.6 Search strategy

The search strategy was devised to capture a broad spectrum of studies, offering insights into the current research landscape. The inclusion criteria encompassed:

a. Original research employing machine learning models for colorectal cancer patient survival prediction.
b. Studies integrating clinical data and gene expression profiles as predictors.
c. Research reporting prediction model accuracy metrics.

d. English language publications from January 2013 to May 2023.

The full electronic search strategy used for the Scopus database for the systematic review of "Machine Learning-Based Prediction of Colorectal Cancer Patient Survival Using Clinical Data and Gene Expression Profiles":

Scopus Database: The search query below on the Scopus database returned 338 document results

TITLE-ABS-KEY(("machine learning"OR"artificial intelligence"OR"predictive modelling"OR ml) AND ("colorectal cancer" OR "colon cancer" OR "rectal cancer") AND ("patient survival" OR "prognosis" OR patient) AND ("clinical data" OR "gene expression" OR "transcriptome" OR "genomics")) AND (LIMIT-TO(PUBYEAR, 2023) OR LIMIT-TO(PUBYEAR, 2022) OR LIMIT-TO(PUBYEAR, 2021) OR LIMIT-TO(PUBYEAR, 2020) OR LIMIT-TO(PUBYEAR, 2019) OR LIMIT-TO(PUBYEAR, 2018) OR LIMIT-TO(PUBYEAR, 2017) OR LIMIT-TO(PUBYEAR, 2016) OR LIMIT-TO(PUBYEAR, 2015) OR LIMIT-TO(PUBYEAR, 2014) OR LIMIT-TO(PUBYEAR, 2013)) AND(LIMIT-TO(PUBSTAGE, "final")) AND(LIMIT-TO(DOCTYPE, "ar")) AND(LIMIT-TO(EXACTKEYWORD, "Human") OR LIMIT-TO(EXACTKEYWORD, "Colorectal Cancer") OR LIMIT-TO(EXACTKEYWORD, "Controlled Study") OR LIMIT-TO(EXACTKEYWORD, "Gene Expression") OR LIMIT-TO(EXACTKEYWORD, "Machine Learning")) AND(LIMIT-TO(LANGUAGE, "English")) AND (LIMIT-TO(SRCTYPE,"j")).

On PubMed Database, the following Search Query returned 145 results

(("Colorectal Neoplasms"[MeSH Terms] OR "colorectal cancer"[All Fields]) AND ("machine learning"[All Fields] OR "artificial intelligence"[All Fields] OR "deep learning"[All Fields] OR "neural network"[All Fields] OR "support vector machine"[All Fields] OR "random forest"[All Fields] OR "decision tree"[All Fields] OR "logistic regression"[All Fields] OR "lasso regression"[All Fields] OR "elastic net regression"[All Fields]) AND ("gene expression"[All Fields] OR "transcriptomics"[All Fields] OR "microarray"[All Fields] OR "RNA-seq"[All Fields] OR "clinical data"[All Fields] OR "demographic data"[All Fields] OR "treatment data"[All Fields] OR "pathological data"[All Fields])) AND (2013:2023[pdat]) **Filters:** Abstract, Full text, Associated data, in the last 10 years, Humans, English, MEDLINE.

2.7 Data management

Search results from Scopus and PubMed were exported in CSV format and subsequently uploaded to Rayyan software for screening. Rayyan.ai software facilitated the efficient screening of the 483 articles that met the eligibility criteria [25]. The final search was conducted on May 13, 2023.

2.8 Study selection

To ensure the inclusion of only the most pertinent and high-quality articles, a rigorous screening process was adhered to, utilizing Rayyan.ai software [26]. Two independent reviewers scrutinized titles and abstracts for relevance. Full-text articles were procured for those aligning with the inclusion criteria or necessitating further evaluation. Discrepancies were addressed through consensus or consultation with a third reviewer [27].

2.9 Data extraction

Two independent reviewers undertook data extraction using a predefined format. Extracted data encompassed study characteristics, patient demographics, sample size, clinical features, gene expression profiles, machine learning techniques, prediction models, and outcome measures. Disagreements were amicably resolved through discussion and, if required, consultation with a third reviewer. The screening trajectory was meticulously documented in the PRISMA flow diagram as captured in Figure 1 [28].

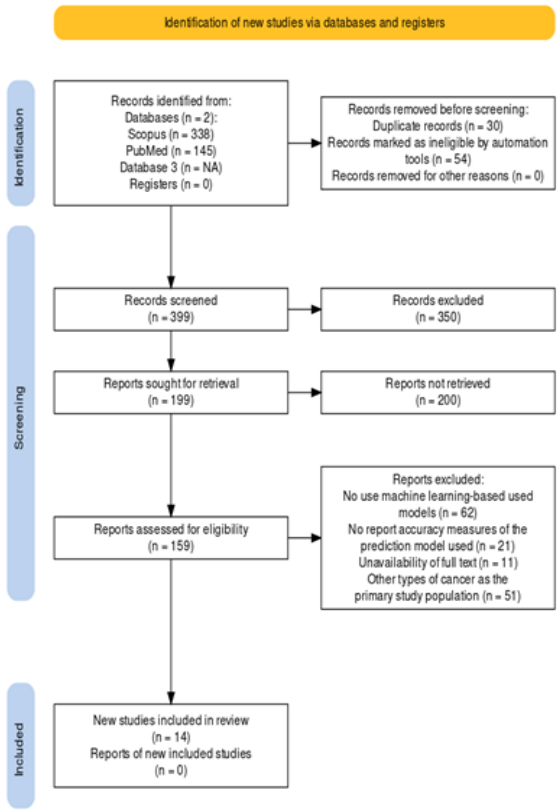

Figure 1. The flow diagram showing the screened studies

2.10 Risk of bias

Bias assessment was meticulously executed, leveraging the researchers' screening method. This involved the application of the eligibility criteria to filter out articles not aligning with the search criteria or not addressing the review's focal topic.

3. RESULTS

In this systematic review, the efficacy of machine learning models in predicting the survival rates of colorectal cancer patients was assessed, focusing on the utilization of clinical data and gene expression profiles. The primary metrics for evaluation were sensitivity, specificity, positive predictive value, negative predictive value, and the area under the curve (AUC). These metrics were not traditional effect measures but served as performance indicators to evaluate the models' predictive capabilities [29-31].

From the initial pool of 483 articles identified, 14 were deemed suitable for the final review after rigorous application of the inclusion and exclusion criteria. This selection process is illustrated in the PRISMA flowchart 2000 [32], as shown in Figure 1.

To ensure the quality of the review, a meticulous process was followed. This involved verifying all duplicates against their sources, analyzing article abstracts in-depth, and evaluating each article against the set inclusion and exclusion criteria. This rigorous screening led to the shortlisting of 9 articles, ensuring the review's quality and relevance.

The characteristics of the 14 finalized studies are summarized in Table 1, highlighting that all were published between 2013 and May 2023.

3.1 Exploration of potential improvements

The machine learning models reviewed demonstrated notable results in predicting colorectal cancer patient survival. However, there is potential for further enhancement. Recent advancements in machine learning, such as deep learning techniques, ensemble methods, and transfer learning, could potentially enhance the models' predictive accuracy. The integration of diverse datasets and real-time patient data could also offer refinements. As the machine learning domain continues to evolve, opportunities to incorporate newer algorithms and techniques will arise, promising further advancements in predictive capabilities.

Table 1. Summary of included studies

S.N	Author(s) (Year)	Title	Summary
1	Wang et al. (2020), [8]	A novel CpG-methylation-based nomogram predicts survival in colorectal cancer	This research developed a nomogram based on CpG-methylation for predicting outcomes in colorectal cancer (CRC). By analyzing methylation data from 378 CRC patients, six significant CpG sites linked to overall survival were identified. This six-CpG marker effectively categorized patients into high and low-risk categories in both training and validation groups. When combined with the TNM stage and age, the nomogram outperformed the three individual prognostic factors in predicting survival. Thus, this CpG-methylation-driven nomogram offers potential as a reliable tool for forecasting CRC patient survival and guiding personalized treatment decisions [8].
2	Liu et al. (2022) [9]	Machine learning-based integration develops an immune-derived lncRNA signature for improving outcomes in colorectal cancer.	This research crafted a consensus signature for colorectal cancer using machine learning, focusing on immune-related long noncoding RNA. Remarkably, this signature surpassed traditional clinical factors, molecular characteristics, and 109 other published signatures in predicting overall survival. It stood out as an independent risk determinant. The team believes that this unique signature holds potential to enhance personalized clinical outcomes for CRC patients [9].
3	Chaba and Omolo (2022) [13]	A machine learning-based approach to cancer classification using RNA-SEQ data	This research assessed four supervised machine-learning methods to categorize colorectal cancer samples based on two clinical endpoints using RNA-Seq data. Drawing from a public colorectal cancer RNA-Seq dataset paired with clinical information, the SVM emerged as the top-performing algorithm, excelling in F-Measure and AUC metrics. It also matched the accuracy of NBLDA. The findings suggest that SVM might be the preferred classification technique for cancer patient data derived from RNA-Seq. The analysis was conducted using the MLSeq package in R [13].
4	Ding et al. (2019) [29]	Predictive biomarkers of colorectal cancer.	This research introduced a computational framework designed to pinpoint biomarkers for colorectal cancer that can be detected in blood, urine, and saliva. This was achieved by merging transcriptomics and proteomics data at a systems biology level. Three distinct models were developed to forecast these biomarkers, and the biological roles and molecular mechanisms of potential biomarkers were deduced. The efficacy of various biomarker combinations was validated using machine learning techniques. ESM1, CTHRC1, and AZGP1 emerged as potent biomarkers for colorectal cancer. Machine learning played a pivotal role in classifying these biomarkers and in analyzing potential targeted treatments, offering valuable insights for clinical care [29].
5	Su et al. (2022) [33]	Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis.	In this research, gene expression data from The Cancer Genome Atlas (TCGA) was harnessed to diagnose and stage colon cancer. The team employed the Weighted Gene Co-expression Network Analysis (WGCNA) and the least absolute shrinkage and selection operator algorithm (Lasso) to delve into differential gene expression and survival patterns. By amalgamating gene modules with Lasso-derived feature genes, they utilized RF, SVM, and decision trees for the diagnosis and staging of colon cancer. Among the models, the RF emerged as the most effective in diagnosing colon cancer and determining its stages. Furthermore, eight genes with significant ties to colon cancer prognosis were pinpointed [33].
6	Kong et al. (2020) [34]	Network-based machine learning in colorectal and bladder organoid models predicts	This research introduces a machine-learning approach designed to pinpoint reliable drug biomarkers for categorizing cancer patients and forecasting their drug reactions. Drawing from network-based evaluations and pharmacogenomic data sourced from three-dimensional organoid culture models, the approach successfully identified biomarkers. These biomarkers

		anti-cancer drug efficacy in patients.	adeptly predicted drug responses in patients with colorectal and bladder cancers. Their accuracy was further corroborated using external transcriptomic datasets and biomarkers based on somatic mutations. The methodology melds gene modules with network-centric strategies, leveraging pharmacogenomic data from organoid models to predict drug responses in cancer patients [34].
7	Yerukala Sathipati et al. (2023) [35]	Artificial intelligence-driven pan-cancer analysis reveals miRNA signatures for cancer stage prediction.	In a quest to pinpoint biomarkers linked to colorectal cancer (CRC) survival, researchers delved into mRNA, miRNA, and tissue microbiome levels. By gathering multi-omics data from CRC samples with short-term (ST) and long-term (LT) survival rates, it was discerned that the CRC tissue microbiome held the most potent predictive capability for three-year patient survival. Notably, distinct microbial communities and gene expressions were observed between the ST and LT groups. This suggests that the bacteria present in CRC tumor tissue could serve as promising biomarkers for forecasting the survival outcomes of CRC patients [35].
8	Yang et al. (2022) [36]	A multi-omics machine learning framework in predicting the survival of colorectal cancer patients.	In an effort to pinpoint biomarkers that could predict colorectal cancer (CRC) survival, this study delved into multi-omics data. Through bioinformatics analysis of 31 short-term survival (ST) and 47 long-term survival (LT) CRC samples, differences in expressed mRNAs and miRNAs were identified, and bacterial community structures were compared. Remarkably, the CRC tissue microbiome emerged as the most potent predictor of three-year survival. The ST group showed a higher abundance of bacteria like Thermoanaerobacterium, Parabacteroides, Oceanicaulis, and Acetonema. In contrast, the LT group was enriched with Methylotenera, Candidatus_Riesia, and Aquamicrobium. By harnessing both bioinformatics and machine learning, the study sheds light on potential biomarkers that could guide personalized therapy strategies for CRC patients [36].
9	Salvucci et al. (2017) [37]	A stepwise integrated approach to personalized risk predictions in stage III colorectal cancer	This research assessed the potential of APOPTO-CELL as a predictive signature for stage III colorectal cancer patients. By analyzing protein concentrations of Procaspase-3, Procaspase-9, SMAC, and XIAP, and employing a machine learning technique using Random Forest, an enhanced signature was identified. Notably, the APOPTO-CELL-PC3 signature outperformed other features in terms of prognostic value and offered more detailed stratification for patients within the CMS1-3 molecular subtype. The findings underscore the value of merging systems-biology-based biomarkers related to apoptosis competency with machine learning methodologies, paving the way for more nuanced patient categorization in clinical care [37].
10	Dai, et al. (2022) [38]	A Novel Pyroptosis-Associated Gene Signature to Predict Prognosis in Patients with Colorectal Cancer	In this study, researchers employed Gene Set Enrichment Analysis to pinpoint pyroptosis-related genes. Utilizing data from the Cancer Genome Atlas and Gene Expression Omnibus databases, they crafted a gene signature to predict the prognosis of colorectal cancer. A set of 12 pyroptosis-associated genes was identified. Based on this signature, patients were categorized into high-risk and low-risk groups. Notably, the high-risk group exhibited poorer outcomes in terms of overall survival, progression-free survival, and relapse-free survival. Additionally, the pyroptosis risk score was found to correlate with immune cell infiltration. This research unveils a new pyroptosis-centric prognostic signature for colorectal cancer, shedding light on its connection with immune infiltration and offering an immunological angle for tailoring personalized treatments [38].
11	Lu et al. (2021) [39]	A 13-immune gene set a signature for the prediction of colon cancer prognosis	In this research, the single sample enrichment analysis (ssGSEA) method was employed to pinpoint an immune gene-set signature with the potential to forecast patient survival and unveil novel therapeutic targets for colon cancer. A signature comprising 13 immune-related genes was formulated and subsequently validated across both training and test datasets. This signature emerged as an independent factor in predicting colon cancer prognosis. Notably, high-risk samples exhibited signs of immunosuppression. Furthermore, the riskScore derived from this study outperformed predictions made by previously published models [39].
12	Zhang et al. (2022) [40]	Bioinformatics analysis reveals immune prognostic markers for overall survival of colorectal cancer patients: a novel machine learning survival predictive system	In a 2022 publication, a team of researchers introduced a machine learning-based system designed to forecast the survival rates of colorectal cancer (CRC) patients using immune gene expression data. By analyzing the differential expression between healthy and tumor tissues, the system pinpointed crucial prognostic immune genes and transcription factors, leading to the creation of an immune-related regulatory network. Utilizing three distinct machine learning algorithms, the system formulated a prognostic model that identified twenty unique risk factors associated with CRC. Impressively, this model could differentiate between patients at high and low risk, offering a valuable resource for tailoring individualized treatment strategies [40].

13	Fu et al. (2023) [41]	Establishment of matrix metalloproteinase 3 time-resolved immunoassay and some potential clinical applications	In this research, the objective was to design a time-resolved fluorescence immunoassay (TRFIA) specifically for the detection of serum matrix metalloproteinase-3 (MMP-3) and to evaluate its clinical relevance in colorectal cancer (CRC) patients. The results demonstrated that the crafted MMP-3 TRFIA exhibited commendable sensitivity, precision, specificity, and recovery rates. Notably, CRC patients exhibited markedly elevated serum MMP-3 levels compared to their healthy counterparts, with this elevation being particularly pronounced in patients with metastatic conditions. These findings suggest a potential link between serum MMP-3 levels and the invasive and metastatic tendencies of CRC. Consequently, serum MMP-3 emerges as a promising auxiliary diagnostic marker for CRC [41].
14	Li et al. (2020) [42]	A multicenter random forest model for effective prognosis prediction in the collaborative clinical research network	In this research, a novel multicenter random forest prognosis prediction model is introduced, tailored for mining clinical data from horizontally partitioned datasets, with a specific emphasis on colorectal cancer. This innovative model addresses and overcomes the performance constraints tied to ensuring data privacy. One of its standout features is its ability to rank feature importance across multiple institutions without the need to consolidate data at a central location. Comparative results indicate that this model not only adheres to privacy-preserving guidelines but also outperforms centrally trained RF models and other potential models in terms of both discrimination and calibration capabilities. By presenting a feasible solution for constructing a prognosis prediction model within a collaborative clinical research framework, this study paves the way for addressing real-world challenges in the application of medical artificial intelligence [42].

3.2 Findings

The 14 studies that met the inclusion criteria collectively analyzed 3,219 colorectal cancer patients. The machine learning techniques used across these studies included decision trees, neural networks, support vector machines, and random forests [31]. The nature of clinical data and gene expression profiles varied among the studies. The reported accuracy measures, primarily sensitivity, specificity, and AUC, indicated AUC values ranging from 0.5 to 0.8, suggesting moderate to high accuracy of the models.

4. DISCUSSION

The systematic review underscores the potential and challenges of utilizing machine learning models in predicting the survival outcomes of colorectal cancer patients. When dissecting the results, it becomes evident that the accuracy of these models is contingent on several factors.

Firstly, the choice of machine learning techniques plays a pivotal role. For instance, traditional techniques like decision trees or support vector machines might offer different predictive capabilities compared to more advanced methods like deep neural networks or ensemble methods. The depth, architecture, and hyperparameters of neural networks, or the combination strategies in ensemble methods, can significantly influence the model's performance. Moreover, the way these techniques are applied, including the handling of imbalances in the dataset, feature selection, and optimization strategies, can also sway the results [43, 44].

Secondly, the nature and quality of the clinical data and gene expression profiles are crucial. Datasets with comprehensive clinical parameters, detailed gene expression profiles, and minimal missing values are more likely to enhance the model's predictive power. However, inconsistencies or biases in data collection, preprocessing, or normalization can introduce noise, potentially reducing the model's accuracy.

Furthermore, the integration of clinical data with gene expression profiles presents its own set of challenges. The high dimensionality of gene expression data, coupled with the heterogeneity of clinical data, necessitates sophisticated feature extraction and selection methods. Techniques like principal component analysis, autoencoders, or domain-specific feature selection can be employed to distill relevant information effectively.

It is also worth noting that the studies included in this review span a decade, from 2013 to 2023. Over this period, the field of machine learning has witnessed rapid advancements. Thus, some of the earlier studies might not have had access to the computational resources or the advanced algorithms available in recent years, potentially influencing their outcomes.

In light of these observations, it is evident that while machine learning holds promise in this domain, achieving high accuracy in predicting colorectal cancer patient survival requires a confluence of the right techniques, quality data, and advanced methodologies. The current landscape suggests a need for more research, especially studies that leverage recent advancements in machine learning, to develop prediction models that not only exhibit high accuracy but are also robust and clinically relevant.

5. CONCLUSION

This systematic review underscores the potential of machine learning in predicting survival outcomes for colorectal cancer patients using clinical and gene expression data. However, the variability in study methodologies and outcomes calls for a measured approach to drawing conclusions. As highlighted in the discussion, there is a basis for exploring more advanced learning techniques to enhance predictive accuracy. The discrepancies across studies emphasize the need for further, more standardized research. By harnessing these advanced techniques and solidifying the evidence base, there is an opportunity to significantly improve colorectal cancer patient prognosis and care. This review not only emphasizes the importance of integrating machine learning into colorectal cancer research but also showcases its potential to revolutionize therapeutic strategies and patient outcomes.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers for their valuable suggestions.

REFERENCE

[1] Koppad, S., Basava, A., Nash, K., Gkoutos, G.V., Acharjee, A. (2022). Machine learning-based identification of colon cancer candidate diagnostics genes. Biology (Basel), 11(3): 365. https://doi.org/10.3390/biology11030365

[2] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 68(6): 394-424. https://acsjournals.onlinelibrary.wiley.com/doi/full/10.3322/caac.21492

[3] Meroueh, C., Chen, Z.E. (2023). Artificial intelligence in anatomical pathology: Building a strong foundation for precision medicine. Human Pathology, 132: 31-38. http://doi.org/10.1016/j.humpath.2022.07.008

[4] Zheng, W., Lu, Y., Feng, X., Yang, C., Qiu, L., Deng, H., Xue, Q., Sun, K. (2021). Improving the overall survival prognosis prediction accuracy: A 9-gene signature in CRC patients. Cancer Medicine, 10: 5998-6009. https://doi.org/10.1002/cam4.4104

[5] Dai, S., Ye, Y., Kong, X., Li J, Ding K (2021). A predictive model for early recurrence of colorectal-cancer liver metastases based on clinical parameters. Gastroenterology Reports, 9: 241-251. https://doi.org/10.1093/gastro/goaa092

[6] Cianci, P., Restini, E. (2021). Artificial intelligence in colorectal cancer management. WArtificial Intelligence in Cancer, 2: 79-89. http://doi.org/10.35713/aic.v2.i6.79

[7] Chen, Y.C., Ke, W.C., Chiu, H.W. (2014). Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. Computer Methods and Programs in Biomedicine, 48: 1-7. https://doi.org/10.1016/j.compbiomed.2014.02.006

[8] Wang, X., Wang, D., Liu, J., Feng M, Wu, X. (2020). A novel CpG-methylation-based nomogram predicts survival in colorectal cancer. Epigenetics, 15: 1213-1227. https://doi.org/10.1080/15592294.2020.1762368

[9] Liu, Z., Liu, L., Weng, S., Guo, C., Dang, Q., Xu, H., Wang, L., Lu, T., Zhang, Y., Sun, Z., Han, X. (2022). Machine learning-based integration develops an immune-derived lncRNA signature for improving outcomes in colorectal cancer. Nature Communications, 13(1): 816. https://doi.org/10.1038/s41467-022-28421-6

[10] Nazari, E., Aghemiri, M., Avan, A., Mehrabian, A., Tabesh, H. (2021). Machine learning approaches for classification of colorectal cancer with and without feature selection method on microarray data. Gene Reports, 25: 101419. https://doi.org/10.1016/j.genrep.2020.101419

[11] American Cancer Society. (2020). Colorectal Cancer Facts & Figures 2020-2022. Atlanta: American Cancer Society. https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/colorectal-cancer-facts-and-figures/colorectal-cancer-facts-and-figures-2020-2022.pdf.

[12] Shiraishi, T., Shinto, E., Nearchou, I.P., Tsuda, H., Kajiwara, Y., Einama, T., Caie, P.D., Kishi, Y., Ueno, H. (2020). Prognostic significance of mesothelin expression in colorectal cancer disclosed by area-specific four-point tissue microarrays. Virchows Archiv, 477(3): 409-420. https://doi.org/10.1007/s00428-020-02775-y

[13] Kaur, P.K., Attwal, K.P.S., Singh, H. (2021). Firefly optimization based noise additive privacy-preserving data classification technique to predict chronic kidney disease. Revue d'Intelligence Artificielle, 35(6): 447-456. https://doi.org/10.18280/ria.350602

[14] Bakrania, A., Joshi, N., Zhao, X., Zheng, G., Bhat, M. (2023). Artificial intelligence in liver cancers: Decoding the impact of machine learning models in clinical diagnosis of primary liver cancers and liver cancer metastases. Pharmacological Research, 189: 106706. https://doi.org/10.1016/j.phrs.2023.106706

[15] Ganguli, R., Franklin, J., Yu, X., Lin, A., Lad, R., Heffernan, D.S. (2022). Machine learning models to prognose 30-Day Mortality in Postoperative Disseminated Cancer Patients. Surgical Oncology, 44: 101810. https://doi.org/10.1016/j.suronc.2022.101810

[16] Siegel, R. L., Miller, K. D., Jemal, A. (2023). Cancer statistics, 2023. CA: A Cancer Journal for Clinicians, 73(1): 17-48. https://doi.org/10.3322/caac.21763

[17] Khabsa, M., Elmagarmid, A., Ilyas, I., Hammady, H., Ouzzani, M. (2015). Learning to identify relevant studies for systematic reviews using random forest and external information. Machine Learning, 102: 465-482.

[18] Zuo, Y., Zhong, J., Bai, H., et al. (2022). Genomic and epigenomic profiles distinguish pulmonary enteric adenocarcinoma from lung metastatic colorectal cancer. eBioMedicine, 82: 104165. https://doi.org/10.1016/j.ebiom.2022.104165

[19] Ben Hamida, A., Devanne, M., Weber, J., Truntzer, C., Derangère V., Ghiringhelli F., Forestier G., Wemmert, C. (2022). Weakly supervised learning using attention gates for colon cancer histopathological image segmentation. Artificial Intelligence in Medicine, 133: 102407. https://doi.org/10.1016/j.artmed.2022.102407

[20] Zheng, Q., Yang, L., Zeng, B., Li, J., Guo, K., Liang, Y., Liao, G. (2021). Artificial intelligence performance in detecting tumor metastasis from medical radiology imaging: A systematic review and meta-analysis. EClinicalMedicine, 31: 100669. https://doi.org/10.1016/j.eclinm.2021.100669

[21] Mahoto, N. A., Shaikh, A., Sulaiman, A., Al Reshan, M.S., Rajab, A., Rajab, K. (2023). A machine learning based data modeling for medical diagnosis. Biomedical Signal Processing and Control, 81: 104481. https://doi.org/10.1016/j.bspc.2023.104481

[22] Moher, D., Liberati, A., Tetzlaff, J., Altman, D.G., Group, P. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. PLoS Medicine, 6(7): e1000097. https://doi.org/10.1371/journal.pmed.1000097

[23] Kastrinos, F., Kupfer, S.S., Gupta, S. (2023). Colorectal cancer risk assessment and precision approaches to screening: Brave new world or worlds apart? Gastroenterology, 164(5): 812-827. https://doi.org/10.1053/j.gastro.2023.02.021

[24] Voets, M.M., Veltman, J., Slump, C.H., Siesling, S., Koffijberg, H. (2022). Systematic review of health economic evaluations focused on artificial intelligence in

healthcare: The tortoise and the cheetah. Value in Health, 25(3): 340-349. http://doi.org/10.1016/j.jval.2021.11.1362

[25] Elmagarmid A, Fedorowicz Z, Hammady H, Ilyas I, Khabsa M, Ouzzani M. (2014). Rayyan: A systematic reviews web app for exploring and filtering searches for eligible studies for cochrane reviews. In Evidence-Informed Public Health: Opportunities and Challenges. Abstracts of the 22nd Cochrane Colloquium, pp. 21-26. John Wiley & Sons Hyderabad, India, India.

[26] Ouzzani, M., Hammady, H., Fedorowicz, Z., Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. Systematic Reviews, 5(1): 210. https://doi.org/10.1186/s13643-016-0384-4

[27] Berg, T., Böhmer, J., Nwaru, B.I., Muche-Borowski, C., Katalinic, A., Lampert, T. (2022). Heart failure in childhood cancer survivors—a systematic review protocol. Systematic Reviews, 11(1): 54. https://doi.org/10.1186/s13643-022-01929-0

[28] Sterne, J.A., Savović, J., Page, M.J., et al. (2019). RoB 2: A revised tool for assessing risk of bias in randomised trials. BMJ, 366: l4898. https://doi.org/10.1136/bmj.l4898

[29] Ding, D., Han, S., Zhang, H., Zhang, S., Lv, Y., Liu, J. (2019). Predictive biomarkers of colorectal cancer. Computational Biology and Chemistry, 83: 107106. https://doi.org/10.1016/j.compbiolchem.2019.107106

[30] Ting, W.C., Chang, H.R., Chang, C.C., Cheng, Y.M., Fang, C.L. (2020). Developing a novel machine learning-based classification scheme for predicting SPCs in colorectal cancer survivors. Applied Sciences, 10(4): 1355. https://doi.org/10.3390/app10051355

[31] Meena, J., Hasija, Y. (2022). Application of explainable artificial intelligence in the identification of Squamous Cell Carcinoma biomarkers. Computational Biology and Medicine, 146: 105505. https://doi.org/10.1016/j.compbiomed.2022.105505

[32] Haddaway, N.R., Page, M.J., Pritchard, C.C., McGuinness, L.A. (2022). PRISMA2020: An R package and Shiny app for producing PRISMA 2020-compliant flow diagrams, with interactivity for optimised digital transparency and Open Synthesis. Campbell Systematic Reviews, 18(2): e1230. https://doi.org/10.1002/cl2.1230

[33] Su, Y., Tian, X., Gao, R., Guo, W., Chen, C., Chen, C., Jia, D., Li, H., Lv, X. (2022). Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis. https://doi.org/10.1016/j.compbiomed.2022.105409

[34] Kong, J., Lee, H., Kim, D., Han, S.K., Ha, D., Shin, K., Kim, S. (2020). Network-based machine learning in colorectal and bladder organoid models predicts anti-cancer drug efficacy in patients. Nature Communications, 11: 5485.

[35] Yerukala Sathipati, S., Tsai, M.-J., Shukla, S.K., Ho, S.Y. (2023). Artificial intelligence-driven pan-cancer analysis reveals miRNA signatures for cancer stage prediction. Human Genetics and Genomics Advances, 4(3): 100190. https://doi.org/10.1016/j.xhgg.2023.100190

[36] Yang, M., Yang, H., Ji, L., Hu, X., Tian, G., Wang, B., Yang, J.L. (2022). A multi-omics machine learning framework in predicting the survival of colorectal cancer patients. Computational Biology and Medicine, 146: 105516. https://doi.org/10.1016/j.compbiomed.2022.105516

[37] Salvucci, M., Urstle, M.L., Morgan, C., et al. (2017). A stepwise integrated approach to personalized risk predictions in stage III colorectal cancer. Clinical Cancer Research 23(5): 1084-2016. http://doi.org/10.1158/1078-0432.CCR-16-1084

[38] Dai, J., Chen, S., Bai, Y., Fu, Y., Pan, Y., Ye, L. (2022). A novel pyroptosis-associated gene signature to predict prognosis in patients with colorectal cancer. Evidence-Based Complementary and Alternative Medicine, 2022: 6965308. https://doi.org/10.1155/2022/6965308

[39] Lu, Z., Chen, J., Yan, J., Liu, Q., Li, F., Xiong, W., Lin, S., Yu, K., Liang, J. (2022). A 13-immune gene set a signature for the prediction of colon cancer prognosis. Combinatorial Chemistry & High Throughput Screening, 24(8): 1205-1216. https://doi.org/10.2174/1386207323666200930104744

[40] Zhang, Z., Huang, L., Li, J., Wang, P. (2022). Bioinformatics analysis reveals immune prognostic markers for overall survival of colorectal cancer patients: A novel machine learning survival predictive system. BMC Bioinformatics, 23(1): 124. https://doi.org/10.1186/s12859-022-04657-3

[41] Fu, Y., Wang, X., Chen, X., Hong, J., Qin, Y., Zhou, Z., Zhou, X., Wang, Y., Zhou, J., Fang, H., Liu, P., Huang, B. (2023). Establishment of matrix metalloproteinase 3 time-resolved immunoassay and some potential clinical applications. Analytical Biochemistry, 666: 115072. https://doi.org/10.1016/j.ab.2023.115072

[42] Li, J., Tian, Y., Zhu, Y., Zhou, T., Li, J., Ding, K., Li, J. (2023). A multicenter random forest model for effective prognosis prediction in the collaborative clinical research network. Artificial Intelligence in Medicine. https://www.x-mol.com/paperRedirect/1225311945510375424.

[43] Manjunath, M.C., Palayyan, B.P. (2023). An efficient crop yield prediction framework using hybrid machine learning model. Revue d'Intelligence Artificielle, 37(4): 1057-1067. https://doi.org/10.18280/ria.370428

[44] Patil, S., Bhosale, S. (2023). Improving cardiovascular disease prognosis using outlier detection and hyperparameter optimization of machine learning models. Revue d'Intelligence Artificielle, 37(4): 1169-1180. http://doi.org/10.18280/ria.370429