Multimedia Tools and Applications
https://doi.org/10.1007/s11042-024-19684-w

Advances in AI-based genomic data analysis for cancer survival prediction

Deepali[1,3] · Neelam Goel[1] · Padmavati Khandnor[2]

Received: 28 June 2023 / Revised: 9 June 2024 / Accepted: 10 June 2024

Abstract

Problem Cancer is one of the deadliest diseases prevalent in the world. Survivability, early diagnosis, and accurate prognosis are of utmost importance for the therapeutics and clinical management of cancer patients. To achieve accurate and timely prediction of the survival of cancer patients, several machine-learning models based on genomic data have been proposed but a comprehensive review of recent applications in cancer survival prediction is lacking.

This paper represents a review of the most recent application of machine learning and deep learning on cancer survival prediction, with a particular focus on the use of genomic data. It specifically targets the most prominent cancer types such as breast cancer, glioblastoma, lung cancer, renal cell cancer, and oral cancer.

Methods A systematic review approach is employed to analyze recent studies on machine learning techniques applied to cancer survival prediction. Emphasis is placed on methodologies utilizing genomic data due to its effectiveness in predicting survival outcomes.

Results This review highlights the efficacy of different machine/deep learning-based techniques in predicting survival outcomes for different cancer types with genomic data. It also provides a summary of the contributions made by different research groups, critically examines the associated challenges, and suggests potential areas for further investigation.

Conclusion Machine learning and deep learning techniques, especially those utilizing genomic data, hold significant promise for accurate cancer survival prediction across diverse cancer types. Despite advancements, challenges such as data heterogeneity and model interpretability still persist. Further research is warranted to address these challenges and develop a comprehensive framework for cancer survival prediction applicable to various cancer types. This review lays the foundation for future investigations in the area of cancer research.

Keywords Cancer · TCGA · Clinical data · Survival time

✉ Neelam Goel
erneelam@pu.ac.in

[1] University Institute of Engineering and Technology, Panjab University, Chandigarh 160014, India

[2] Department of Computer Science, Punjab Engineering College (deemed to be university), Chandigarh 160012, India

[3] Department of Computer Science, Guru Nanak College, Budhlada 151502, India

Published online: 21 June 2024

1 Introduction

Cancer is a deadly disease that caused about 10 million deaths worldwide in 2020 and will have the greatest relative increase by 2040 [1]. In the United States, one in ten adults has been infected with cancer disease [2]. These figures are not different for India, having the third rank in cancer cases around the globe [3]. The death rate due to cancer in India has been increasing rapidly since 1990 [4]. These figures are truly astonishing and scary. Usually, Cancer occurs due to the abnormal growth of cells [5]and metastasizes to various parts of the body through blood or lymph vessels [6]. Cancer cells acquire the needed space and nutrients of the healthy organ which then may lead to organ failure and can become a cause of death. Over the past decades, cancer research has been directed toward detecting cancer at its initial stages. The researcher's continuous efforts led to the development of new techniques and strategies for the early prediction of cancer which helps in its treatment [7], like [8] Ostu's thresholding technique for the segmentation of brain tumors using MRI images. Treatment of cancer patients further can be improved by survival time analysis [9]. The survival time refers to the duration between the date of diagnosis or initiation of treatment for a disease, such as cancer, and the conclusion of the observation period. Predicting cancer survivability is challenging due to the complex nature of cancer disease which is influenced by various genetic and environmental factors. Additionally, incomplete or noisy data, limited sample sizes, and variability in data collection protocols pose obstacles to accurate prediction. It's crucial for therapeutics and clinical management because timely and accurate predictions can guide treatment decisions and help improve patient outcomes.

The approach for assessing the effectiveness of a new treatment in a clinical trial is to measure the overall survival. The accurate prediction of survival time can provide doctors with a better approach to the treatment of a person who is suffering from a disease. In past decades, high-throughput technologies have been utilized to predict survival time which helps to define prognostic indices for mortality or recurrence of disease and to thoroughly investigate the outcome of treatment. Survival time is computed using clinical, image, or genomic data. Among the different types of data, genomic data is the most valuable to have an accurate prediction. Genomic data provide information about molecular mechanisms of cancer development and progression. At the same time, other data types like clinical data or imaging data do not capture genetic changes in cancer patients. But including genetic data for research, helps in predicting survival time accurately by exploring the complexities of disease and also enables in identification of biomarkers of survival. These genomic data can be obtained from various open-source platforms like Gene Expression Omnibus (GEO) [10] and The Cancer Genome Atlas (TCGA) [11, 12]. AI and Machine learning are useful in the medical and healthcare fields, including disease detection, healthcare services, and industry applications. This review article concentrates on the application of both conventional machine learning and deep learning methods that utilize genomic data for predicting survival time. As genomic data becomes increasingly standardized and sophisticated analysis techniques continue to evolve, it has the potential to significantly enhance the development of robust algorithms for predicting survival time.

The highly complex as well as expensive genomic data analysis is a significant burden for clinicians in terms of diagnosis, prediction, and subsequent management. Correspondingly, the diagnosis and treatment planning are stagnant and fallible, as these rely on the physician's skills and expertise, which may be instinctive and inaccurate. Hence, quantitative measures are required and are best for the diagnosis. Advanced machine learning and deep learning techniques provide targeted solutions. By employing new learning

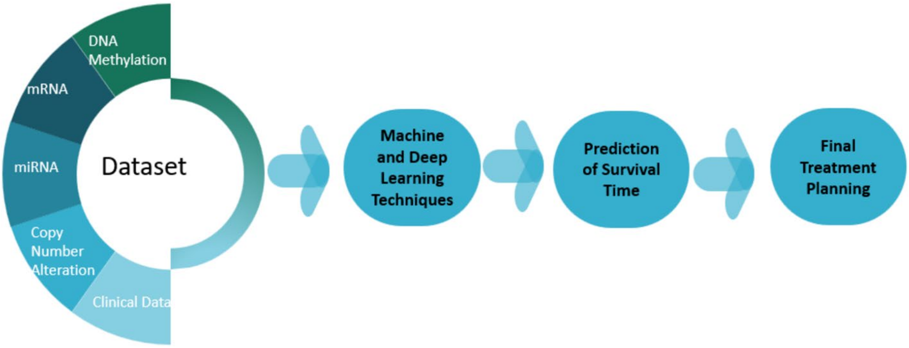

Fig. 1 Use of machine/deep learning techniques for survival prediction of cancer patients

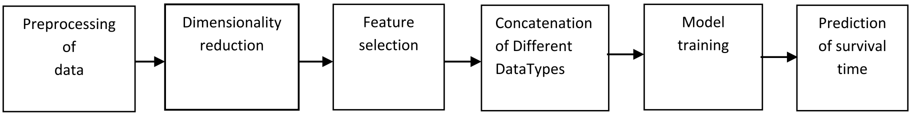

Fig. 2 End-to-end framework for prediction of survival time using machine learning techniques

techniques, clinicians can enhance the treatment planning of patients and achieve better outcomes. The use of machine learning and deep learning techniques in utilizing various types of genomic data for predicting the survival time is shown in Fig. 1.

The prediction of survival time for cancer patients comes with a few sequential steps: (1) preprocessing of genomic data, (2) dimensionality reduction, (3) feature selection, (4) model training, and (5) prediction of survival time. In the training stage, various type of genomic data (DNA Methylation [13], Copy number Alteration, mRNA, and other datatypes) is preprocessed. These features are then employed, either individually or in combination, to reduce the dimensionality, and the model is trained using various machine-learning techniques. Figure 2 illustrates these steps of survival prediction using machine learning techniques.

On the other side, advanced deep learning methods that utilize layered artificial neural networks (ANN) with supervised or unsupervised learning techniques automatically combine feature selection, dimensionality reduction, and prediction into a single process. As deep learning models strive to discover concealed patterns and connections, they usually perform better in predicting survival time than traditional machine learning approaches. With the increasing availability of genomic data and advanced processing capabilities, deep learning is gaining popularity as a powerful tool for genomic data analysis.

Existing literature for predicting survival time in cancer patients often relies on clinical or image data. However, these methods may not always provide accurate predictions, and these do not fully utilize the wealth of information available in genomic data. With genomic data, there is very little review on the survival prediction of cancer patients, but they are not too extensive and also there is no such comparison of different machine learning models which gives the idea of potential research in the future. This paper aims to comprehensively review the published literature regarding the use of machine learning and deep learning techniques for survival prediction of breast, glioblastoma, lung, renal cell, and oral cancer. The primary outcome indicator is the various biomarkers of survival of these cancer types and the accuracy in the prediction of survival using genomic data.

The main contributions of this paper are as follows:

- Comprehensive review of the use of machine learning and deep learning techniques for survival prediction in various types of cancer.
- Analysis of the effectiveness of these techniques in utilizing various types of genomic data for predicting the survival time.
- Analysis of various feature selection methods for extracting the important features for survival prediction.
- Identification of potential areas for future research and improvement in this field.

The rest of this review is organized as follows. Section 2 provides a detailed overview of the machine-learning approaches for cancer survival prediction. Section 3 presents a discussion of these techniques. Finally, Section 4 concludes the review and provides directions for future research.

2 Machine learning approaches for cancer survival prediction

2.1 Public databases for genomic dataset

Datasets are extensively accessible and selected based on the discretion of various research groups. However, most of them opt to build algorithms utilizing well-established cancer patient databases to enhance the research value. The most well-known databases for genomic data are The Cancer Genome Atlas (TCGA) and the International Cancer Genome Consortium (ICGC). TCGA is the largest database with a dataset of more than 33 types of cancer that is freely available (https://portal.gdc.cancer.gov/). Another well-established database is The International Cancer Genome Consortium (ICGC) (https://dcc.icgc.org/) giving cancer researchers access to over 2,400 consistently analyzed genomes corresponding to over 1,100 unique ICGC donors.

2.2 Performance measurements

In Cancer Research, different research groups utilize various algorithms with the use of different types of datasets. For comparison, various metrics are used by researchers. Some commonly used metrics are c-index, accuracy, sensitivity, specificity, mean mean squared error or mean absolute deviation, precision and recall, and overall survival time.

2.3 Selection procedure

The research was performed on the IEEE Xplore digital library, Web of Science, Science Direct, Google Scholar, and PubMed search engines using keywords “Genomic data”, “Survival” with “Breast cancer”, “Glioblastoma”, “Lung cancer”, “Renal cell cancer”, and “Oral cancer”. Afterward, these keywords are combined using the “AND” operator with “Machine Learning” or “Deep Learning” with “Survival prediction”. The articles that are in journals or conference papers were included. The inclusion criteria of the articles are machine learning or deep learning models in the prediction of survival using genomic data. The most recent papers were included in the study.

2.4 Machine learning approaches for survival prediction of cancer

Clinically, the survival prediction ability of cancer patients has a great impact as it helps in planning the treatment of patients [14]. This could be advantageous not only for the patients themselves but also for their families, who often undergo significant stress when dealing with cancer patients. In the long term, it could also result in cost savings for treatment. Therefore, in cancer research, much research focuses on predicting the survival time of patients and on predicting survival time using imaging datasets, clinical datasets, and genomic datasets [15]. As genomic data increases the accuracy in the prediction of survival time, it is worth including this genomic data for the study. However, with the inclusion of genomic data, huge dimensionality becomes a major problem that needs to be considered. Thus, several researchers make use of dimensionality reduction techniques to deal with the problem of high dimensionality. The survival prediction model has two main components: dimensionality reduction or feature selection and prediction model. These techniques are mainly of two types namely supervised and unsupervised. A few popular feature selection or dimensionality reduction techniques are Principal Component Analysis (PCA) [16], Non-negative factorization (NMF) [17], and Factor Analysis [16], etc. With feature selection or dimensionality reduction methods, the feature can be reduced. Additionally, recent advancements in digital healthcare have focused on utilizing fog and cloud networks for cancer detection, introducing novel paradigms such as the Multi-Cancer Multi-Omics Clinical Dataset Laboratories (MCMOCL) Schemes which incorporate federated learning, auto-encoder, and XGBoost methods to improve accuracy, reduce processing delay, and enhance security in heterogeneous cancer clinics [18] and other studies have explored hybrid cancer detection schemes utilizing SARSA reinforcement learning and multi-omics data processing in fog cloud networks, aiming to enhance accuracy and reduce processing time in distributed clinical settings [19].

Machine learning techniques help in extracting meaningful patterns from complex genomic data which is not feasible by traditional analytical methods [20]. By incorporating machine learning techniques with genomic data, the survival time can be predicted accurately. Several machine learning techniques can be used for the prediction and some commonly used techniques are Support Vector Machine (SVM), AdaBoost, Decision trees, and Random forest. While developing machine learning techniques for survival time prediction using genomic data, research should consider various factors such as feature selection, model interpretability, and validation methodologies. Feature selection techniques aim to identify the most informative genomic features while reducing noise and overfitting. Model interpretability ensures that predictions are clinically actionable. Validation methodologies such as cross-validation and external validation assess model generalizability and robustness across diverse datasets.

There are various challenges in applying machine learning to genomic data analysis for cancer survival prediction including data heterogeneity, incomplete data, feature biases, model overfitting, and interpretability issues. Limited sample sizes and imbalanced datasets can lead to biased model performance and poor generalizability. Furthermore, the complexity of genomic data necessitates sophisticated feature engineering and regularization techniques to prevent overfitting and enhance model interpretability. Addressing these challenges is critical to ensuring the reliability and clinical utility of predictive models.

Recent research on survival prediction using genomic data has focused on various cancer types such as breast cancer, lung cancer, colorectal cancer, ovarian cancer, glioblastoma, oral cancer, renal cell cancer, and cervical cancer. But this study focused on breast cancer, lung cancer, renal cell cancer, oral cancer, and glioblastoma. These cancer types represent a diverse spectrum of diseases with distinct molecular characteristics and prognostic factors, making them ideal candidates for genomic data analysis to improve survival prediction accuracy.

Some researchers focused on finding the biomarkers of the survival time of cancer using various machine learning algorithms. Various studies revealed that different Long non-coding RNA (lncRNA) [15, 21–23], genomic instability derived lncRNA [24], and autophagy-associated long noncoding RNAs (ARlncRNAs) [25, 26] act as a biomarker for the prediction of survival time. Apart from lncRNA, the Fanconi anemia pathway can act as a prognostic biomarker for survival prediction [27]. Studies also demonstratedthat transfer learning-based deep features [28],radiomics signature [29], ten glucose metabolism risk signature [30], prognostic index, stem cell-related gene signature [31], seven CPG-based signature [32], 6-gene signature [33], aggregated signature based on ligand-gated channel pathways [34], Radiomics signature [35], TLS [36], APE1 Polymorphism [37], Tp53 [28], COL4A5, ABCB1, NR3C2 and PLG [26] can act as predictive biomarkers while 5-snoRNA signature [38], cancer-associated fibroblasts [39], TP53 [40]are not a promising predictor for survival prediction. Research has been also performed to show the importance of tumor environment [41], age [42], and oral hygiene [43] in the prediction of survival time. Some authors identified various miRNA or mRNA genes [44, 45]and nomogram-based genes or miRNA signatures [45–49] which act as predictors for survival. Another study [50]implemented the ESTIMATE machine learning algorithm which identified IL10, IGLL5, and POU2AF1 prognostic biomarkers. The research was performed for lung adenocarcinoma risk by using Random forest, Univariate Cox, and SigFeature algorithms which identified 16-gene expression having a high correlation with patient risk [51]. Another study showed the significance of somatic mutation in survival prediction [52]. A study on the effect of race on survival concluded that African Americans had prolonged survival [53] and another approach interpreted that the difference in survival rates depends on gender [54]. Identifying biomarkers associated with survival time is a key focus of research, but challenges persist in determining the most predictive feature among the vast genomic features.

Probing further, researchers compared the different machine learning methods in the survival prediction. For instance, the survival time was estimated by implementing various algorithms namely: 1-Nearest Neighbor (1NN), Naive Bayes (NB), SVM, AdaBoost, Tree Random Forest (TRF), Radial Basis Function Network (RBFN), and Multilayer Perceptron models, out of which Trees Random Forest model (TRF) which is a rule-based classification model turns out to be the best in prediction with the highest level of precision [55]. Furthermore, a study that used six different machine learning models AdaBoost, NB, SVM, RF, Adabag, Least-Squares SVM (LSSVM), and two classical methods Logistic Regression (LR) and Linear Discriminant Analysis (LDA), for predicting survival time and metastasis of breast cancer concluded that SVM outshined other models by providing more accurate data [56].

Some researchers proposed new models for survival prediction using genomic data. In an experiment, Genomic data, and Pathological images Multiple Kernel Learning (GPMKL) model, based on Multiple Kernel Learning (MKL) used integrated pathological images and genomic data for survival time prediction. This model was created to execute feature fusion, which is a crucial aspect of breast cancer classification. The results indicate that integrating genomic data with pathological images produces better outcomes than using either genomic data or pathological images alone, for GPMKL with 95% specificity, sensitivity, accuracy, and precision were increased by 4.3%, 0.9%, and 3.8% respectively as compared to Genomic data based Multiple Kernel Learning (GMKL) and improved by 13.9%, 3.2%, and 16.4% as compared to Pathological images based Multiple Kernel Learning (PMKL) which proved GPMKL as deserving and useful in predicting human breast cancer survival [57].In another study, Immunohistochemistry was used to cluster the dataset based on receptor status in which significant variables were ranked by the random forest variable selection method and there was a multiplatform network named Multimodal AutoEncoders (MAE) that was implemented to classify breast cancer patients based on their survival rates and their subtypes. Survival rate prediction was performed using multitype modalities and the lowest mean square error was achieved with gene expression (0.16541). Moreover, decision tree (DT), NB, K-Nearest Neighbor (KNN), LR, SVM, RF, and gradient boosting trees (GBT) were implemented for the survival prediction which concluded that GBT and RF-based classifiers or regression models performed best [58]. A new algorithm Crystall was proposed for breast cancer which predicted the survival time of patients and classified the patients based on their survival time that is whether a patient would live longer than 5 years or not. The proposed one performed better for both problems and achieved a mean absolute error of 31.62 days for predicting how long a breast cancer patient will live within 5 years [59]. To improve the disease-free survival prediction performance of lung squamous cell carcinoma, a novel method named LSCDFS-MKL was proposed, which is based on multiple kernel learning. The model used the Gradient descent algorithm for solving various kernel learning problems and integrated pathological images and genomic data. The method increased the specificity, accuracy, and sensitivity by 2.20%, 2.68%, and 7.14% than to using genomic data only and improved by 9.89%, 24.11%, and 34.02% compared with pathological images only. The accuracy of LSCDFS-MKL was 100% for the prediction of disease-free survival and performed better than other prediction methods [60].

A new Ordinal Multi-Modal Feature Selection (OMMFS) framework [61] was developed to identify the features from pathological images, DNA methylation, mRNA, and copy number variation and used a sparse canonical correlational analysis framework with ordinal survival information. The results showed that this method has a better performance in patient stratification and can be used as the general framework for any cancer type for the prediction of biomarkers or to predict the response of any treatment. Another stratification method based on the Elastic net penalized Cox proportional hazard regression model was designed to group the advanced-stage oral cancer patients into different risk groups using genetic and clinicopathological features [62], which helped create an online calculator.

A novel integrative model based on the Bayesian averaging model for renal cell carcinoma was proposed, which used the dimensionality reduction technique PCA and Sparse PCA (SPCA) to generate features of the low dimension of three genomic data types and considered the interaction between the data types. The mean square error was calculated for both dimensionality reduction techniques and compared the results with and without the consideration of the interaction between data types. Results showed that the mean square error was the least for PCA with interaction (2.07). These models also validated the ccRCC-based biomarkers for renal cell carcinoma which was verified in the literature [63]. Another novel machine learning model [64]employed coherent voting networks and predicted the survival time of breast cancer accurately.

Other research [69] used the SVM model to investigate the relationship between glioma topographic location and molecular characteristics and suggested that tumor location plays a role in glioma development and could be used to improve treatment and predict outcomes. Another study [70] explored the role of anoikis resistance in breast cancer metastasis and treatment optimization. Through a comprehensive analysis of mRNA and lncRNA profiles, ten key mRNAs and six lncRNAs associated with anoikis were identified using the LASSO Cox regression model. [71] integrated multi-omics data with clinical factors to identify significant biomarkers for Glioblastoma Multiforme (GBM) prognosis. Employing the Multimodal iterative Random Forest (MiRF) algorithm, 35 molecular features comprising 19 genes and 16 proteins were isolated, distinguishing between short-term and long-term survival as well as high and low Karnofsky performance scores. Another study [72] integrated histology and genomics using the probabilistic graphical model framework. It used the multilayer perceptron model to generate informative embeddings capturing underlying cancer properties by canonical correlation analysis (CCA) and penalized variants (pCCA), and the model generates informative. Other research [73] utilized multi-omics data from lung adenocarcinoma patients to improve survival prediction accuracy. By using novel feature extraction techniques and unbiased selection methods, 32 molecular features were identified from the TCGA dataset, achieving an AUC of 0.839 for a 2-year survival prediction model.

Several machine learning algorithms have been implemented. However, the choice of algorithm often depends on specific cancer types and dataset characteristics.

Conventional machine learning models used in the prediction of survival of breast, glioblastoma, lung, renal cell, and oral cancer using genomic data are presented in Table 1. Machine learning algorithms predict cancer patients' survival very efficiently but need time-consuming and complex pre-processing techniques. The problems faced by machine learning models can be minimized with the use of deep learning techniques.

2.5 Deep learning approaches for survival prediction of cancer

Deep Learning is an advanced method of machine learning for the processing of complex data. The process is regarded as deep because it comprises hidden layers, where the output of one layer is passed to the next. In contrast to traditional machine learning, deep learning algorithms typically do not require prior feature selection or extensive data pre-processing (although some pre-processing may be necessary). Instead, they employ either supervised or unsupervised training with multiple layers. In the existing literature, there were some review papers like [74] that review the various deep learning models based on multi-omics data for clinical implications and their challenges in using multi-omics data.

A deep learning method named Multimodal Deep Neural [75] was applied for survival prediction using clinical data, copy number alteration, and integrating gene expression. Three deep neural networks were constructed by taking into account the different data types to create a multimodal network. An Attention-based MultiNonnegative Matrix Factorization (AMND) algorithm was designed by integrating gene expression and clinical data. Nonnegative Matrix Factorization algorithms [75] were used to compute eigenvector weights to extract useful information from gene expression and clinical data. After that, the summation of weights of eigenvectors was concatenated with clinical data to feed into deep neural networks for classification. Deep learning-based concatenation autoencoder (ConcatAE) was developed to integrate features of different datasets and used cross-modality autoencoder (CrossAE) which predicts the overall survival time [77]. A new autoencoder-based feature extraction method named DeepSGP [78]was presented for glioblastoma patients which stratified the patients into

Table 1 Summary of machine learning methods for survival prediction of cancer patients using genomic data

[Reference] Year	Country of research	Data Source	No. of Samples	Techniques used for Missing Value	Pre-processing Described	Type of Cancer	Technique	Evaluation Parameters
[15]2016	China	TCGA Database	1064	No	No	Breast Cancer	Cox Regression	Specificity, Sensitivity, Accuracy
[44]2017	China	TCGA Database	521	No	No	Lung Cancer	Cox Regression	Survival rate
[27]2018	USA	TCGA Database	1091	No	No	Breast Cancer	Cox Regression	C-index
[65]2018	China	TCGA Database	1100	Weighted nearest neighbors algorithm	Z-score normalization	Breast Cancer	GPMKL	Sensitivity, Specificity, Precision, Accuracy, Mattew's correlational coefficient
[66]2018	South Korea	TCGA Database	868	Median imputation	Normalized for the mean to be 0 and standard deviation to be 1	Breast Cancer	Random Walk	Accuracy
[42]2018	Canada	TCGA, Cancer Genomics of Kidney Databases	525	No	No	Renal Cell Cancer	Cox Regression	False discovery rate, Survival rate
[34]2018	United States	TCGA Database	264	Yes	preprocessed with RSEM software	Oral Cancer	Monte Carlo cross-validation	Sensitivity, Specificity
[47]2018	USA	TCGA Database	541	Yes	No	Oral Cancer	Random Forest	Survival Probability
[23]2019	China	TCGA Database	1016	No	No	Breast Cancer	Cox Regression	Sensitivity, Specificity, Overall Survival

Table 1 (continued)

[Reference] Year	Country of research	Data Source	No. of Samples	Techniques used for Missing Value	Pre-processing Described	Type of Cancer	Technique	Evaluation Parameters
[58]2019	Germany	TCGA Database	1046	No	No	Breast Cancer	Multimodal Autoencoder	Mean square error, Precision, Recall, Mattew's correlational coefficient, Accuracy
[33] 2019	China	TCGA and CGGA Databases	295	No	No	Glioblastoma	Cox Regression	Sensitivity, Specificity, Overall Survival
[35] 2019	China	TCGA	147	No	No	Glioblastoma	Multivariate Cox regression	C-index
[53]2019	USA	Northwestern Medicine Enterprise Data Warehouse (NMEDW) and TCGA Database	995	predictive mean matching (PMM)	No	Glioblastoma	Cox Proportional	C-index
[60]2019	China	TCGA Database	101	Weighted nearest neighbors algorithm	Z-score normalization	Lung Cancer	Multiple Kernel Method	Accuracy, Precision, Specificity, Sensitivity, Mattew's correlational coefficient
[51]2019	China	TCGA Database	492	No	Robust Multichip Average (RMA) algorithm	Lung Cancer	LASSO Cox Model	Overall Survival, Survival probability, Sensitivity, Specificity, Accuracy
[63]2019	India	Synapse	243	No	No	Renal Cell Cancer	Bayesian averaging model	Mean Square error

Table 1 (continued)

[Reference] Year	Country of research	Data Source	No. of Samples	Techniques used for Missing Value	Pre-processing Described	Type of Cancer	Technique	Evaluation Parameters
[37]2019	Taiwan	Kaohsiung Medical university	451	No	No	Oral Cancer	Cox proportional hazard	Survival rate
[40]2019	Canada	London Health Sciences Center	136	No	No	Oral Cancer	Cox proportional hazard regression	Survival probability
[43]2019	Taiwan	National Cheng Kung University	740	No	No	Oral Cancer	Cox proportional hazard	Overall Survival
[67]2020	UK	TCGA Database	955	No	Robust Multichip Average (RMA) algorithm	Glioblastoma	Cox Regression	Sensitivity, Specificity, Overall Survival
[68]2020	China	TCGA Database	402	No	No	Lung Cancer	Cox Regression	C-index
[50]2020	China	TCGA Database	537	No	No	Renal Cell Cancer	ESTIMATE	Survival rate
[32] 2020	China	TCGA Database	485	No	upper quartile normalization	Renal Cell Cancer	Multivariate Cox Regression	Overall Survival, Specificity, Risk Score, survival probability, Sensitivity, Accuracy
[46] 2020	China	TCGA Database, ArrayExPrecisionss	333	No	No	Renal Cell Cancer	LASSO Cox Regression	Specificity, Overall Survival
[63]2019	India	Synapse	243	No	No	Renal Cell Cancer	Bayesian averaging model	Mean Square error
[37]2019	Taiwan	Kaohsiung Medical university	451	No	No	Oral Cancer	Cox proportional hazard	Survival rate

Table 1 (continued)

[Reference] Year	Country of research	Data Source	No. of Samples	Techniques used for Missing Value	Pre-processing Described	Type of Cancer	Technique	Evaluation Parameters
[40]2019	Canada	London Health Sciences Center	136	No	No	Oral Cancer	Cox proportional hazard regression	Survival probability
[43]2019	Taiwan	National Cheng Kung University	740	No	No	Oral Cancer	Cox proportional hazard	Overall Survival
[67]2020	UK	TCGA Database	955	No	Robust Multi-chip Average (RMA) algorithm	Glioblastoma	Cox Regression	Sensitivity, Specificity, Overall Survival
[68]2020	China	TCGA Database	402	No	No	Lung Cancer	Cox Regression	C-index
[50]2020	China	TCGA Database	537	No	No	Renal Cell Cancer	ESTIMATE	Survival rate
[32] 2020	China	TCGA Database	485	No	upper quartile normalization	Renal Cell Cancer	Multivariate Cox Regression	Overall Survival, Specificity, Risk Score, survival probability, Sensitivity, Accuracy
[46] 2020	China	TCGA Database, ArrayExPrecisionss	333	No	No	Renal Cell Cancer	LASSO Cox Regression	Specificity, Overall Survival
[52] 2020	China	TCGA	488	No	log2 transformed and RSEM-normalized and Z-score normalization	Breast Cancer	MKL	Precision, Sensitivity, Specificity, Accuracy
[21] 2021	China	UCSC Xena Database, ICGC	1284	No	No	Breast Cancer	Cox and LASSO regression	Overall Survival, Sensitivity, Specificity

Table 1 (continued)

[Reference] Year	Country of research	Data Source	No. of Samples	Techniques used for Missing Value	Pre-processing Described	Type of Cancer	Technique	Evaluation Parameters
[59] 2021	China	TCGA Database	928	No	Filtering out top-ranked 100,000 features with the largest variances	Breast Cancer	Crystal Algorithm	Accuracy, Sensitivity, Specificity, Mean Absolute Error
[22] 2021	China	UCSC Xena Browser, CGGA	1722	No	Data normalization	Glioblastoma	Cox proportional hazard	Survival probability, True positive rate, False positive rate
[30] 2021	China	Repository for Molecular Brain Neoplasia Data, TCGA, and CGGA	696	No	No	Glioblastoma	Cox Regression	Survival probability, True positive rate, False positive rate
[31] 2021	China	TCGA Database	497	Removing samples with missing value	No	Lung Cancer	multivariate Cox regression model	Sensitivity, Survival probability
[41] 2021	China	TCGA Database	576	no	pre-processed by the limma algorithm	Lung Cancer	Cox Regression	Sensitivity, Specificity, Percent Survival
[45] 2021	China	TCGA Database	309	No	No	Renal Cell Cancer	Cox Regression	Overall Survival
[64] 2021	Italy	Metabric Consortium	2000	missing data handling methodologies that rely on interpolation	Log2 transformation normalization	Breast Cancer	Coherent Voting Network	Overall Survival, Disease-free survival

Table 1 (continued)

[Reference] Year	Country of research	Data Source	No. of Samples	Techniques used for Missing Value	Pre-processing Described	Type of Cancer	Technique	Evaluation Parameters
[39] 2021	Korea	TCGA Database	1053	No	No	Lung Cancer	gradient boosting machine algorithm	Sensitivity, Specificity, Disease-free survival
[24] 2021	Italy	TCGA Database	539	No	No	Renal Cell Cancer	LASSO Cox Regression	Survival Rate, True positive rate, False positive rate
[25] 2021	China	TCGA Database	530	No	No	Renal Cell Cancer	Cox Regression	C-index
[26] 2021	China	TCGA Database	537	No	No	Renal Cell Cancer	Cox Regression	C-index
[69] 2022	China	Harbin Medical University	586	No	Yes	Glioblastoma	SVM	Accuracy
[70] 2023	China	TCGA Database	113	Yes	No	Breast Cancer	LASSO	C-index
[71] 2023	Arabia	TCGA Database, GDAC Firehose,cBioportal	596	Yes	No	Glioblastoma	MiRF	C-index
[72] 2024	USA	TCGA Database	1095	No	No	Breast Cancer	Multilayer Perceptron Model	Correlational
[73] 2024	Poland	TCGA Database, CPTAC-3	363	Yes	No	Lung Cancer	LASSO	AUC

different survival groups with an accuracy of 0.83 and a prediction of overall survival with an accuracy of 0.89. Another artificial intelligence-based approach [79]was proposed which used the SVM model on radiomic features and Cox-PH regression model with radiomics signature, clinical and genomics data to categorize the patients into different risk groups which gave the c-index of 0.75 with the combination of the dataset and 0.65 for clinical data only. In a different deep learning model, deep orthogonal fusion [80]used multimodal data to predict the overall survival of glioblastoma patients with a c-index of 0.788. DeepSurv with multi-omics data [81] for oral cancer predicted survival time with a c-index of 0.94. A deep learning approach was used to analyze the tumor-infiltrating lymphocyte (TIL) profiles to identify the association with survival. Also, 16 out of 22 TILs were different for predicted risk groups [82]2.

The study [83] utilized machine learning and deep learning techniques to identify prognostic biomarkers for predicting the time-to-development of oral cancer and stratifying survival among patients with premalignant lesions. Autoencoder deep learning neural network extracts features, which were further analyzed using a univariate Cox regression model. Supervised clustering based on encoded features distinguished high-risk and low-risk groups, while a random forest classifier identified gene profiles associated with oral cancer subtypes. Another research [84] introduced a hybrid deep learning model with clinical, gene expression, and copy alteration data for breast cancer prediction and survival prediction of patients. A novel predictive model [85] using a graph convolutional network (GCN) and Choquet fuzzy ensemble, integrating multi-omics and clinical data was introduced. The model achieved competitive performance metrics, including an accuracy of 0.820 and a balanced accuracy of 0.769, outperforming baseline models and demonstrating its efficacy in prognostic classification. A novel prognostic algorithm [86] by integrating pathogenomics and AI-based techniques. Machine learning and deep learning algorithms identified predictive features for survival outcomes, with the multimodal which outperforms unimodal and suggesting potential for personalized treatment strategies in oral cancer. A study introduced a Deep Convolution Cascade Attention Fusion Network (DCCAFN) for predicting lung cancer patients' survival based on imaging genomics [87]. The DCCAFN demonstrated effectiveness in multimodal data fusion, aiding physicians in risk stratification and personalized treatment decisions to improve patient's quality of life.

These studies show the potential of deep learning in enhancing survival prediction accuracy across various cancer types. Table 2 shows the deep learning algorithms for survival prediction of breast, glioblastoma, lung, renal cell, and oral cancer using genomic data.

The comparison of various machine and deep learning algorithms for survival prediction in terms of accuracy evaluation parameters are shown in Table 3 which can help the researchers to select the best algorithm for survival prediction.

2.6 Feature selection methods for survival prediction of cancer

Dimensionality reduction techniques can help reduce the number of features in a dataset by identifying a smaller set of representative features that capture the most important information. However, even after dimensionality reduction, there may still be redundant or irrelevant features in the remaining set of features. The researchers have used feature selection methods that can help address this issue by identifying and selecting only the most relevant features for a particular task. The following are the methods that are commonly used by researchers for dimensionality reduction or feature selection techniques.

Table 2 Summary of deep learning methods for survival prediction of cancer patients using genomic data

[Reference] Year	Country of research	Data Source	No. of Samples	Techniques used for Missing Value	Pre-processing Described	Type of Cancer	Technique	Evaluation Parameters
[75] 2019	China	Metabric Consortium	1980	weighted nearest neighbors algorithm	min–max normalization	Breast Cancer	Multimodal Deep Neural network	Sensitivity, Specificity, Precision, Accuracy, Mattew's correlational coefficient
[76] 2019	China	Metabric Consortium	1980	No	No	Breast Cancer	Attention-based Multi NMF Deep Neural Network	Precision, Recall, Accuracy
[77] 2020	USA	Modified National Institute of Standards and Technology (MNIST) database and TCGA Databases	1060	Remove features with missing data	log2 transformation and min–max normalization	Breast Cancer	ConcatAE and CrossAE	Precision, Recall, Accuracy
[78] 2021	India	TCGA Databases	129	No	EdgeR package for data preprocessing	Glioblastoma	DeepSGP	Accuracy
[80] 2021	USA	TCGA Databases	176	No	No	Glioblastoma	Deep Orthogonal Fusion	C-index
[82] 2021	Korea	TCGA Database	566	No	CIBERSORT	Oral Cancer	DNN	Survival, Accuracy
[79] 2022	USA	Hospital of the University of Pennsylvania (HUP)	516	No	Z-score normalization removed all features with small variations with mean absolute deviation	Glioblastoma	SVM, Cox-PH Regression	C-index

Table 2 (continued)

[Reference] Year	Country of research	Data Source	No. of Samples	Techniques used for Missing Value	Pre-processing Described	Type of Cancer	Technique	Evaluation Parameters
[81]2022	India	TCGA Database	500	Yes	Min–Max Scalar normalization, encode categorical to numerical values	Oral Cancer	DeepSurv	C-index
[83] 2023	Iran	GEO	86	No	Yes	Oral Cancer	Autoencoder Deep Learning Neural Network	Accuracy
[84] 2023	Egypt	METABRIC dataset	128	Yes	Yes	Breast Cancer	Convolutuonal Neural Network	Accuracy
[85] 2023	India	METABRIC dataset	1980	Yes	Yes	Breast Cancer	Choquet Fuzzy Ensemble	Accuracy, Sensitivity, Specificity, F1-measure
[86] 2024	Germany	TCGA Database	406	No	Yes	Oral Cancer	Random Survival Forest, Gradient Boosting Survival Analysis, Cox PH, Fast Survival SVM, and DeepSurv	C-index
[87] 2024	China	TCGA Database	168	Yes	Yes	Lung Cancer	DCCAFN	Accuracy, Precision, Recall, F1-measure

Table 3 Comparison of various machine and deep learning methods for survival prediction of cancer patients using genomic data in terms of accuracy

Reference	Model	Accuracy	Type of Cancer
[15]	Cox Regression	0.714	Breast Cancer
[25]	Cox Regression	0.809	Renal Cell Cancer
[28]	DeepSurv	0.81	Oral Cancer
	Cox proportional hazard model	0.756	
	Random Forest	0.77	
[32]	Multivariate Cox Regression	0.821	Renal Cell Cancer
[38]	LASSO	0.704	Oral Cancer
[49]	LASSO with nomogram	0.792	Renal Cell Cancer
[51]	LASSO	0.753	Lung Cancer
[52]	Multiple Kernel Learning	0.9808	Breast Cancer
[55]	Naïve Bayes	0.95	Breast Cancer
	1-Nearest Neighbors	0.91	
	Support Vector Machine	0.94	
	Tree Random Forest	0.96	
	AdaBoost	0.94	
	Multilayer Perceptron	0.95	
	RBF Network	0.95	
[56]	Logistic Regression	0.93	Breast Cancer
	Linear Discriminant Analysis	0.93	
	Support Vector Machine	0.92	
	Naïve Bayes	0.92	
	Random Forest	0.92	
	AdaBoost	0.89	
	Least-square SVM	0.91	
[57]	GPMKL	0.86	Breast Cancer
[58]	Multimodal Autoencoders	0.91	Breast Cancer
[59]	Crystall	0.9276	Breast Cancer
[60]	LSCDFS-MKL	0.8022	Lung Cancer
[66]	Autoencoders	0.81	Breast Cancer
[69]	SVM	0.71	Glioblastoma
[75]	Multimodal Deep Neural Network by integrating Multi-dimensional Data	0.826	Breast Cancer
	Logistic Regression	0.76	
	Random Forest	0.791	
	Support Vector Machine	0.805	
[76]	Attention-Based Multi-NMF Deep Neural Network	0.848	Breast Cancer
[77]	ConcatAE	0.962	Breast Cancer
	CrossAE	0.963	
[78]	DeepSGP	0.89	Glioblastoma
[82]	CIBERSORT	0.972	Oral Cancer
[83]	Autoencoder Deep Learning Neural Network	0.916	Oral Cancer
[84]	Convolutuonal Neural Network	0.98	Breast Cancer
[85]	Choquet Fuzzy Ensemble	0.82	Breast Cancer
[87]	DCCAFN	0.831	Lung Cancer

I. **Factor Analysis:-** Factor analysis is a dimensionality reduction technique that simplifies complex data sets by identifying underlying factors or dimensions that explain the patterns and relationships in the data. It identifies key factors that contribute to the variance in the data [88].

II. **Principal Component Analysis (PCA):-**PCA is a widely used dimensionality reduction technique that identifies the key features or components that explain the variance in a dataset. It works by transforming the original variables into a new set of uncorrelated variables called principal components, which capture the most important information in the data [89].

III. **Sparse PCA:-**Sparse PCA is a variant of PCA that produces sparse solutions by promoting sparsity in the loadings of the principal components. This means that it identifies a smaller number of key features or components that contribute most to the variance in the data while setting the remaining loadings to zero [90].

IV. **Kernel PCA:-**Kernel PCA is a nonlinear dimensionality reduction technique that extends the linear PCA to handle nonlinear relationships in the data. It works by projecting the data into a high-dimensional feature space using a nonlinear kernel function and then applying PCA to the resulting kernel matrix. This allows it to capture nonlinear variations in the data and identify the key components that explain the variance in the feature space [91].

V. **LASSO:-**LASSO (Least Absolute Shrinkage and Selection Operator) is a dimensionality reduction technique that selects a subset of relevant features by imposing a penalty on the absolute values of the regression coefficients. This encourages sparsity in the model and effectively sets some of the coefficients to zero, leading to a simpler and more interpretable model [92].

VI. **Autoencoder:-**Autoencoder is a neural network architecture that can be used for unsupervised dimensionality reduction. It works by encoding the input data into a lower-dimensional representation, also known as a latent space, and then decoding it back to the original dimensions. The encoder and decoder are trained together to minimize the reconstruction error between the input and output data. By constraining the size of the latent space, the autoencoder can effectively reduce the dimensionality of the input data, while preserving its essential features [93].

VII. **Fselector:-** In Fselector feature selection, the F-test is used to measure the dependence between each feature and the target variable. The F-test calculates a score for each feature, which represents the degree of correlation between the feature and the target variable [94].

VIII. **mRMR:-** The mRMR algorithm selects features by maximizing the relevance criterion and minimizing the redundancy criterion. It first selects the feature with the highest relevance and then selects additional features that have high relevance but low redundancy with the previously selected features. This process continues until the desired number of features is selected [95].

IX. **Non-negative Matrix Factorization (NMF):-** The NMF algorithm decomposes a given matrix X into two non-negative matrices W and H, where W represents the set of basis vectors (latent features) and H represents the set of coefficients (weights) that combine these basis vectors to approximate the original matrix X. The NMF algorithm seeks to find the best values for W and H such that their product approximates the original matrix X [96].

X. **Log-rank test:-** The Log-rank test works by dividing the population into two or more groups based on the values of a given feature. It then calculates the survival function for each group and compares them using a statistical test such as the log-rank test or the Wilcoxon test. The p-value obtained from the test indicates whether the survival curves are significantly different or not. If the p-value is below a certain threshold (e.g., 0.05), it suggests that the feature is an important predictor of survival [97].

XI. **Minimal Depth:**- The algorithm works by constructing a decision tree using all available features and calculating the minimal depth of each feature. The features with the smallest minimal depth are considered the most important, as they appear closer to the root of the decision tree and have a greater influence on the final decision [98].

XII. **Linear Correlation:**- The algorithm works by calculating the Pearson correlation coefficient between each feature and the target variable. The Pearson correlation coefficient measures the linear relationship between two variables, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation), with 0 indicating no correlation. Features with a high absolute value of the Pearson correlation coefficient are considered strongly correlated with the target variable and are selected for further analysis [99].

XIII. **Transfer Learning** learning is a feature selection algorithm used in machine learning to identify the most relevant features for a target task by leveraging knowledge from a related source task. The algorithm works by first training a model on a related source task using a large set of features. The trained model is then used to extract features from the source task that are relevant to the target task. These extracted features are then used as the input for a model trained on the target task [100].

XIV. **Scree Plot:**- Scree plot feature selection is a graphical method used in the principal component analysis (PCA) to identify the most important principal components (PCs) and, consequently, the most relevant features in a dataset. To use scree plot feature selection, the number of principal components to retain is selected based on the elbow in the scree plot. The corresponding PCs and their corresponding loadings (weights) are then used as the most important features in the dataset [101].

The comparison of various machine learning models based on the dimensionality reduction or feature selection method is shown in Table 4. Despite the advancements in feature selection, challenges remain in identifying the most informative features for survival prediction. The selection of appropriate feature selection methods depends on specific datasets and cancer types representing an ongoing research gap in the field.

Overall, machine learning and deep learning approaches hold promise for enhancing cancer survival prediction, and addressing research gaps related to feature selection, algorithm selection, and model interpretability is essential for advancing the field and translating findings into clinical practice.

3 Discussion

Machine learning techniques have shown promising results by improving the accuracy of survival time prediction of cancer patients. By integrating multiple types of omics data, such as DNA methylation, copy number alteration, and mRNA expression, machine learning algorithms can identify patterns and relationships that may not be apparent through individual omics analyses. Numerous studies have investigated the identification of biomarkers for predicting the survival time of cancer patients using various machine-learning algorithms. Different Long non-coding RNA (lncRNA) [12, 16–21], the Fanconi anemia pathway [22], transfer learning-based deep features [23], radiomics signature [24, 30], ten

Table 4 Comparison of models based on feature selection and dimensionality reduction methods

References	Dimensionality Reduction Techniques					Feature Selection Methods								
	Factor Analysis	**PCA**	**Sparse PCA**	**Kernel PCA**	**LASSO**	**Fselector**	**Autoencoder**	**mRMR**	**NMF**	**Log Rank Test**	**Minimal Depth**	**Linear Correlation**	**Transfer Learning**	**Scree Plot**
[23]	×	×	×	×	×	×	×	×	×	×	×	×	√	×
[33]	×	×	×	×	√	×	×	×	×	×	×	×	×	×
[49]	×	×	×	×	×	×	×	√	×	×	×	×	×	×
[60]	×	×	×	×	×	×	×	×	×	×	×	√	×	×
[61]	×	×	×	×	×	×	×	×	×	√	×	×	×	×
[58]	×	√	√	×	×	×	×	×	×	×	×	×	×	×
[65]	×	×	×	×	×	√	×	×	×	×	×	×	×	×
[66]	√	×	×	×	×	×	√	×	×	×	×	×	×	×
[68]	×	×	×	×	√	×	×	×	×	×	×	×	×	×
[75]	×	×	×	×	×	×	×	√	×	×	×	×	×	×
[76]	×	√	×	√	×	×	√	×	×	√	√	×	×	×
[78]	×	×	×	×	×	×	√	×	×	×	×	×	×	×

glucose metabolism risk signature [25], prognostic index, stem cell-related gene signature [26], seven CPG-based signature [27], 6-gene signature [28], aggregated signature based on ligand-gated channel pathways [29], TLS [31], APE1 Polymorphism [32], Tp53 [23, 35], COL4A5, ABCB1, NR3C2 and PLG [21], and ESTIMATE machine learning algorithm [47] have been found to act as prognostic biomarkers. In contrast, 5-snoRNA signature [33], cancer-associated fibroblasts [34], and TP53 [35] have not shown promising results for survival prediction. Additionally, the effect of the tumor environment [36], age [37], and oral hygiene [38] on survival time prediction has been investigated. Some authors have identified various miRNA or mRNA genes [39–42] and nomogram-based genes or miRNA signatures [40, 43–46] that act as predictors for survival. Several studies have compared different machine learning methods, including 1-Nearest Neighbor (1NN), Naive Bayes (NB), Support Vector Machine (SVM), AdaBoost, Tree Random Forest (TRF), Radial Basis Function Network (RBFN), Multilayer Perceptron, AdaBoost, Least-Squares SVM (LSSVM), Logistic Regression (LR), and Linear Discriminant Analysis (LDA) [52, 53]. The TRF and SVM models have been found to provide the best results in predicting survival time and metastasis of breast cancer [52, 53]. Some researchers have proposed new models for survival prediction using genomic data, including GPMKL based on Multiple Kernel Learning (MKL) [54], Multimodal AutoEncoders (MAE) [55], and Crystall [56]. These models have demonstrated improved accuracy and precision in predicting human breast cancer and lung squamous cell carcinoma survival time.

A variety of deep learning-based methods have been applied to predict survival in cancer patients using multimodal data, including clinical data, copy number alteration, gene expression, and radiomic features. These methods include Multimodal Deep Neural [62], Attention-based MultiNonnegative Matrix Factorization (AMND) [62], ConcatAE [77], CrossAE [77], DeepSGP [66], SVM model [67], Cox-PH regression model [67], deep orthogonal fusion [68], DeepSurv [75], and TIL profiling [76]. These methods have demonstrated high accuracy in predicting overall survival, with c-index values ranging from 0.75 to 0.94.

To the best of our knowledge, this is the first review of the application of Machine Learning to survival prediction by making use of genomic data for breast, glioblastoma, lung, renal cell, and oral cancer. In the review, most of the studies used the open-access database. However, there are certain issues with public databases like data is not updated at regular intervals. Therefore, research should focus on collecting data from different private or public hospitals by obtaining ethical consent from patients and hospitals.

Various feature selection methods used by researchers are Fselector [65], autoencoder [58, 78],mRMR [75], NMF [76], Long Rank test [61], minimal depth [65], DeepSGP [78], Transfer learning [23], Linear correlation [60] and ScreePlot [75]. There is a need to compare the various feature selection methods in predicting the survival time of cancer for cancer. This would help the researchers to choose the best feature selection method.

The size of the dataset of cancer patients used in the current study is between 100 to 2000 and genomic data has a large number of features that are difficult to process with machine learning algorithms. Therefore, there is a need to use appropriate feature selection or dimensionality reduction techniques to select important features. Also, the dataset contains outliers, noise, and missing values. In future studies, researchers should not only use appropriate machine learning methods but also consider various preprocessing and feature selection methods.

The most commonly used algorithms in this review are Cox regression, LASSO regression, Random forest, and Machine kernel learning, and only a few studies used the deep learning approaches. In future studies, deep learning approaches should be explored for the survival prediction of cancer patients using genomic data. Different techniques can be combined to produce the best results.

There are various areas where further investigation or improvement can be performed such as comparing the results by applying both early and late integration for multi-omics data which can lead to selecting the best integration approach for survival prediction. It has further scope to consider intra and inter-interaction effects between different data types of multi-omics data. There is a need to compare various feature selection techniques to identify the most effective methods for predicting cancer survival. This would help researchers choose the best approach based on the specific characteristics of their dataset. While traditional machine learning methods like Cox regression, LASSO regression, Random forest, and Machine kernel learning have been commonly used, only a few studies have explored deep learning approaches. There is scope for applying deep learning approaches for survival prediction using multi-omics data. Further, machine learning and deep learning techniques can be combined to achieve the best performance which can be further explored.

4 Conclusion

The paper is a comprehensive review of the most recent machine learning-based approaches for predicting cancer patient survival, with a focus on the use of genomic data. The paper covers various cancer types, including breast cancer, glioblastoma, lung cancer, renal cell cancer, and oral cancer, and discusses the use of different machine learning techniques, such as random forests, support vector machines, neural networks, and deep learning algorithms. This paper also highlights the challenges involved in developing accurate survival prediction models, such as the need for large and standardized datasets with detailed genomic and clinical information. In addition, various dimensionality and feature selection methods are also compared, which can help improve the accuracy and generalizability of the models.

The key contribution of the research is to highlight the impact of machine and deep learning in the survival prediction of cancer patients. This review paper helps researchers explore the potential of an integrative approach to genomic data in survival prediction, which helps clinicians make informed decisions that further improve treatment outcomes.

Despite the advancements highlighted in this review, several limitations persist in the field of cancer survival prediction. One notable limitation is the access to standardized data, which may introduce biases in the prediction. Deep learning and machine learning started a new revolution in the survival prediction of cancer patients and there is still much scope for further improvement. Data from cancer patients have different formats, e.g., miRNA, mRNA, copy number variation, clinical data, etc. With the advancement of new technologies, working with these data types has become easy but there is still a need to explore various feature selection or dimensionality reduction techniques for handling a large number of features of genomic data. Addressing these limitations will be crucial for realizing the full potential of machine learning in survival prediction.

In the coming times, work should continue focusing on testing and improving the algorithm and state-of-the-art models to improve cancer patients' survival prediction. Moreover, there is a great scope to work with time-series data of cancer patients for better prognosis and to improve survival time. The impact of early and late integration of genomic data on survival prediction can further be explored.

Funding Not Applicable.

Availability of data and material Not Applicable.

Code availability Not Applicable.

Declarations

Conflict of Interest Authors declare that there is no conflict of interest.

References

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A et al (2021) Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 71:209–249. https://doi.org/10.3322/caac.21660
2. Siegel RL, Miller KD, Jemal A (2019) Cancer statistics, 2019. CA Cancer J Clin 69:7–34. https://doi.org/10.3322/caac.21551
3. World Cancer Day (2020): Facts about the deadly disease killing one person every 8 minutes - SCIENCE News n.d. https://www.indiatoday.in/science/story/world-can-day-2019-cancer-causes-cures-treatments-myths-1446568-2019-02-04 Accessed 4 July 2020
4. Smith RD, Mallath MK. History of the Growing Burden of Cancer in India: From Antiquity to the 21st Century. J Glob Oncol 2019:1–15. https://doi.org/10.1200/jgo.19.00048
5. Hanahan D, Weinberg RA (2000) The Hallmarks of Cancer Review evolve progressively from normalcy via a series of pre. Cell 100:57–70
6. Kashyap D, Garg VK, Goel N. Intrinsic and extrinsic pathways of apoptosis: Role in cancer development and prognosis. Adv Protein Chem Struct Biol, vol. 125, Academic Press Inc.; 2021, p. 73–120. https://doi.org/10.1016/bs.apcsb.2021.01.003
7. Kourou K, Exarchos TP, Exarchos KP, Karamouzis M v., Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J 2015;13:8–17. https://doi.org/10.1016/j.csbj.2014.11.005
8. Nyo MT, Mebarek-Oudina F, Hlaing SS, Khan NA (2022) Otsu's thresholding technique for MRI image brain tumor segmentation. Multimedia tools and applications 81(30):43837–43849
9. Blay JY, Penel N, Valentin T, Anract P, Duffaud F, Dufresne A, Verret B, Cordoba A, Italiano A, Brahmi M, Henon C (2024) Improved nationwide survival of sarcoma patients with a network of reference centers. Ann Oncol 35(4):351–363. https://doi.org/10.1016/j.annonc.2024.01.001
10. Zhang X, Zhang W, Jiang Y, Liu K, Ran L, Song F (2019) Identification of functional lncRNAs in gastric cancer by integrative analysis of GEO and TCGA data. J Cell Biochem 120:17898–17911. https://doi.org/10.1002/jcb.29058
11. Tomczak K, Czerwińska P, Wiznerowicz M (2015) The Cancer Genome Atlas (TCGA): An immeasurable source of knowledge. Wspolczesna Onkologia 1A:A68-77. https://doi.org/10.5114/wo.2014.47136
12. Deepali, Goel N, Khandnor P. TCGA: A multi-genomics material repository for cancer research. Mater Today Proc, vol. 28, Elsevier Ltd; 2020, p. 1492–5. https://doi.org/10.1016/j.matpr.2020.04.827
13. Goel N, Karir P, Garg VK (2017) Role of DNA methylation in human age prediction. Mech Ageing Dev 166:33–41. https://doi.org/10.1016/j.mad.2017.08.012
14. Sharma D, Goel N, Kumar V (2022) Predicting Survivability in Oral Cancer Patients. Proceedings of 2nd International Conference on Artificial Intelligence: Advances and Applications: ICAIAA 2021. Springer Nature Singapore, Singapore, pp 153–62
15. Guo W, Wang Q, Zhan Y, Chen X, Yu Q, Zhang J et al (2016) Transcriptome sequencing uncovers a three–long noncoding RNA signature in predicting breast cancer survival. Sci Rep 6:1–10. https://doi.org/10.1038/srep27931
16. Usman Ali M, Ahmed S, Ferzund J, Mehmood A, Rehman A (2017) Using PCA and factor analysis for dimensionality reduction of bio-informatics data. Int J Adv Comput Sci Appl (IJACSA) 8(5):415-426 https://doi.org/10.14569/IJACSA.2017.080551
17. Jing-Yan Wang J, Wang X, Gao X (2013) Non-negative matrix factorization by maximizing correntropy for cancer clustering. BMC Bioinformatics 14:107–117

18. Mohammed MA, Lakhan A, Abdulkareem KH, Garcia-Zapirain B (2023) Federated auto-encoder and XGBoost schemes for multi-omics cancer detection in distributed fog computing paradigm. Chemom Intell Lab Syst 15(241):104932
19. Mohammed MA, Lakhan A, Abdulkareem KH, Garcia-Zapirain B (2023) A hybrid cancer prediction based on multi-omics data and reinforcement learning state action reward state action (SARSA). Comput Biol Med 1(154):106617
20. Ali AM, Mohammed MA (2024) A Comprehensive Review of Artificial Intelligence Approaches in Omics Data Processing: Evaluating Progress and Challenges. International Journal of Mathematics, Statistics, and Computer Science 2:114–167
21. Huang Z, Xiao C, Zhang F, Zhou Z, Yu L, Ye C, et al. A Novel Framework to Predict Breast Cancer Prognosis Using Immune-Associated LncRNAs. Front Genet 2021;11. https://doi.org/10.3389/fgene.2020.634195
22. Liu G, Liu D, Huang J, Li J, Wang C, Liu G, et al. Comprehensive analysis of ceRNA network related to lincRNA in glioblastoma and prediction of clinical prognosis. BMC Cancer 2021;21. https://doi.org/10.1186/s12885-021-07817-5
23. Sui Y, Shao B (2019) A lymph node metastasis-related protein-coding genes combining with long noncoding RNA signature for breast cancer survival prediction. Cellular Physiology 234:20036–20045. https://doi.org/10.1002/jcp.28600
24. Yang H, Xiong X, Li H. Development and Interpretation of a Genomic Instability Derived lncRNAs Based Risk Signature as a Predictor of Prognosis for Clear Cell Renal Cell Carcinoma Patients. Front Oncol 2021;11. https://doi.org/10.3389/fonc.2021.678253
25. Yu JJ, Mao WP, Xu B, Chen M (2021) Construction and validation of an autophagy-related long noncoding RNA signature for prognosis prediction in kidney renal clear cell carcinoma patients. Cancer Med 10:2359–2369. https://doi.org/10.1002/cam4.3820
26. Xuan Y, Chen W, Liu K, Gao Y, Zuo S, Wang B, et al. A Risk Signature with Autophagy-Related Long Noncoding RNAs for Predicting the Prognosis of Clear Cell Renal Cell Carcinoma: Based on the TCGA Database and Bioinformatics. Dis Markers 2021;2021. https://doi.org/10.1155/2021/8849977
27. Zhang Y, Yang W, Li D, Yang JY, Guan R, Yang MQ. Toward the precision breast cancer survival prediction utilizing combined whole genome-wide expression and somatic mutation analysis. BMC Med Genomics 2018;11:99–107. https://doi.org/10.1186/s12920-018-0419-x
28. Kim DW, Lee S, Kwon S, Nam W, Cha IH, Kim HJ (2019) Deep learning-based survival prediction of oral cancer patients. Sci Rep 9:1–10. https://doi.org/10.1038/s41598-019-43372-7
29. Jovcevska I. Genetic secrets of long-term glioblastoma survivors. Bosn J Basic Med Sci 2019;19:116–24. https://doi.org/10.17305/bjbms.2018.3717.
30. Zhang C, Wang M, Ji F, Peng Y, Wang B, Zhao J, et al. A Novel Glucose Metabolism-Related Gene Signature for Overall Survival Prediction in Patients with Glioblastoma. Biomed Res Int 2021;2021. https://doi.org/10.1155/2021/8872977
31. Huang Z, Shi M, Zhou H, Wang J, Zhang HJ, Shi JH. Prognostic signature of lung adenocarcinoma based on stem cell-related genes. Sci Rep 2021;11. https://doi.org/10.1038/s41598-020-80453-4
32. Xu L, He J, Cai Q, Li M, Pu X, Guo Y (2020) An effective seven-CpG-based signature to predict survival in renal clear cell carcinoma by integrating DNA methylation and gene expression. Life Sci 243:117289. https://doi.org/10.1016/j.lfs.2020.117289
33. Zuo S, Zhang X, Wang L (2019) An RNA sequencing-based six-gene signature for survival prediction in patients with glioblastoma. Sci Rep 9:1–10. https://doi.org/10.1038/s41598-019-39273-4
34. Schomberg J, Ziogas A, Anton-Culver H, Norden-Krichmar T (2018) Identification of a gene expression signature predicting survival in oral cavity squamous cell carcinoma using Monte Carlo cross-validation. Oral Oncol 78:72–79. https://doi.org/10.1016/j.oraloncology.2018.01.012
35. Tan Y, Mu W, Wang X chun, Yang G qiang, Gillies RJ, Zhang H. Improving survival prediction of high-grade glioma via machine learning techniques based on MRI radiomic, genetic and clinical risk factors. Eur J Radiol 2019;120:108609. https://doi.org/10.1016/j.ejrad.2019.07.010
36. Li K, Guo Q, Zhang X, Dong X, Liu W, Zhang A et al (2020) Oral cancer-associated tertiary lymphoid structures: gene expression profile and prognostic value. Clin Exp Immunol 199:172–181. https://doi.org/10.1111/cei.13389
37. Huang HI, Chen CH, Wang SH, Wang LH, Lin YC (2019) Effects of APE1 Asp148Glu polymorphisms on OPMD malignant transformation, and on susceptibility to and overall survival of oral cancer in Taiwan. Head Neck 41:1557–1564. https://doi.org/10.1002/hed.25576
38. Xing L (2020) Expression scoring of a small-nucleolar-RNA signature identified by machine learning serves as a prognostic predictor for head and neck cancer. J Cell Physiol 235:8071–8084. https://doi.org/10.1002/jcp.29462
39. Min KW, Kim DH, Noh YK, Son BK, Kwon MJ, Moon JY. Sci Rep 2021;11. https://doi.org/10.1038/s41598-021-96344-1

40. Mundi N, Prokopec SD, Ghasemi F, Warner A, Patel K, MacNeil D, et al. Genomic and human papillomavirus profiling of an oral cancer cohort identifies TP53 as a predictor of overall survival. Cancers Head Neck 2019;4. https://doi.org/10.1186/s41199-019-0045-0
41. Chen J, Zhou R. Tumor microenvironment related novel signature predict lung adenocarcinoma survival. PeerJ 2021;9. https://doi.org/10.7717/peerj.10628
42. Feulner L, Najafabadi HS, Tanguay S, Rak J, Riazalhosseini Y (2019) Age-related variations in gene expression patterns of renal cell carcinoma. Urologic Oncology: Seminars and Original Investigations 37:166–175. https://doi.org/10.1016/j.urolonc.2018.11.006
43. Chang CC, Lee WT, Hsiao JR, Ou CY, Huang CC, Tsai ST et al (2019) Oral hygiene and the overall survival of head and neck cancer patients. Cancer Med 8:1854–1864. https://doi.org/10.1002/cam4.2059
44. Li X, An Z, Li P, Liu H (2017) A predictive model for lung adenocarcinoma patient survival with a focus on four miRNAs. Oncol Lett 14:2991–2995. https://doi.org/10.3892/ol.2017.6481
45. Zhou J, Liu G, Wu X, Zhou Z, Li J, Ji Z. A Risk Score Model Based on Nine Differentially Methylated mRNAs for Predicting Prognosis of Patients with Clear Cell Renal Cell Carcinoma. Dis Markers 2021;2021. https://doi.org/10.1155/2021/8863799.
46. Zhang C, D M, Wang F, D M, Guo F, D M, et al. A 13-gene risk score system and a nomogram survival model for predicting the prognosis of clear cell renal cell carcinoma. Urol Oncol 2020;38:74.e1–74.e11. https://doi.org/10.1016/j.urolonc.2019.12.022
47. Nunez Lopez YO, Victoria B, Golusinski P, Golusinski W, Masternak MM (2018) Characteristic miRNA expression signature and random forest survival analysis identify potential cancer-driving miRNAs in a broad range of head and neck squamous cell carcinoma subtypes. Reports of Practical Oncology and Radiotherapy 23:6–20. https://doi.org/10.1016/j.rpor.2017.10.003
48. Zhang Z, Lin E, Zhuang H, Xie L, Feng X, Liu J, et al. Construction of a novel gene-based model for prognosis prediction of clear cell renal cell carcinoma. Cancer Cell Int 2020;20. https://doi.org/10.1186/s12935-020-1113-6
49. Zhao E, Bai X. Nomogram Based on microRNA Signature Contributes to Improve Survival Prediction of Clear Cell Renal Cell Carcinoma. Biomed Res Int 2020;2020. https://doi.org/10.1155/2020/7434737
50. Zeng Q, Zhang W, Li X, Lai J, Li Z. Bioinformatic identification of renal cell carcinoma microenvironment-associated biomarkers with therapeutic and prognostic value. Life Sci 2020;243. https://doi.org/10.1016/j.lfs.2020.117273
51. Li Y, Ge D, Gu J, Xu F, Zhu Q, Lu C (2019) A large cohort study identifying a novel prognosis prediction model for lung adenocarcinoma through machine learning strategies. BMC Cancer 19:886. https://doi.org/10.1186/s12885-019-6101-7
52. He Z, Zhang J, Yuan X, Zhang Y. Integrating Somatic Mutations for Breast Cancer Survival Prediction Using Machine Learning Methods. Front Genet 2021;11. https://doi.org/10.3389/fgene.2020.632901
53. Wu M, Miska J, Xiao T, Zhang P, Kane JR, Balyasnikova I v., et al. Race influences survival in glioblastoma patients with KPS $\geq$ 80 and is associated with genetic markers of retinoic acid metabolism. J Neurooncol 2019;142:375–84. https://doi.org/10.1007/s11060-019-03110-5
54. Daripally S, Peddi K. Polymorphic variants of drug-metabolizing enzymes alter the risk and survival of oral cancer patients. 3 Biotech 2020;10. https://doi.org/10.1007/s13205-020-02526-5
55. Montazeri M, Montazeri M, Montazeri M, Beigzadeh A (2016) Machine learning models in breast cancer survival prediction. Technol Health Care 24:31–42. https://doi.org/10.3233/THC-151071
56. Tapak L, Shirmohammadi-Khorram N, Amini P, Alafchi B, Hamidi O, Poorolajal J (2019) Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. Clin Epidemiol Glob Health 7:293–299. https://doi.org/10.1016/j.cegh.2018.10.003
57. Sun D, Li A, Tang B, Wang M. Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome. Comput Methods Programs Biomed 2018;161:45–53. https://doi.org/10.1016/j.cmpb.2018.04.008
58. Karim MR, Wicaksono G, Costa IG, Decker S, Beyan O (2019) Prognostically Relevant Subtypes and Survival Prediction for Breast Cancer Based on Multimodal Genomics Data. IEEE Access 7:133850–133864. https://doi.org/10.1109/ACCESS.2019.2941796
59. Liu S, Li H, Zheng Q, Yang L, Duan M, Feng X et al (2021) Survival Time Prediction of Breast Cancer Patients Using Feature Selection Algorithm Crystall. IEEE Access 9:24433–24445. https://doi.org/10.1109/ACCESS.2021.3054823
60. Zhang A, Li A, He J, Wang M. LSCDFS-MKL: A multiple kernel-based method for lung squamous cell carcinomas disease-free survival prediction with pathological and genomic data. J Biomed Inform 2019;94. https://doi.org/10.1016/j.jbi.2019.103194

61. Shao W, Han Z, Cheng J, Cheng L, Wang T, Sun L et al (2020) Integrative Analysis of Pathological Images and Multi-Dimensional Genomic Data for Early-Stage Cancer Prognosis. IEEE Trans Med Imaging 39:99–110. https://doi.org/10.1109/TMI.2019.2920608
62. Tseng YJ, Wang HY, Lin TW, Lu JJ, Hsieh CH, Liao CT. Development of a Machine Learning Model for Survival Risk Stratification of Patients with Advanced Oral Cancer. JAMA Netw Open 2020;3. https://doi.org/10.1001/jamanetworkopen.2020.11768
63. Singh A, Goel N, Yogita. Integrative Analysis of Multi-Genomic Data for Kidney Renal Cell Carcinoma. Interdiscip Sci 2020;12:12–23. https://doi.org/10.1007/s12539-019-00345-8
64. Pellegrini M. Accurate prediction of breast cancer survival through coherent voting networks with gene expression profiling. Sci Rep 2021;11. https://doi.org/10.1038/s41598-021-94243-z
65. Sun D, Li A, Tang B, Wang M (2018) Integrating genomic data and pathological images to effectively predict breast cancer clinical outcome. Comput Methods Programs Biomed 161:45–53. https://doi.org/10.1016/j.cmpb.2018.04.008
66. Kim SY, Kim TR, Jeong H, Sohn K. Integrative pathway-based survival prediction utilizing the interaction between gene expression and DNA methylation in breast cancer. BMC Med Genomics 2018;11. https://doi.org/10.1186/s12920-018-0389-z
67. Prasad B, Tian Y, Li X (2020) Large-Scale Analysis Reveals Gene Signature for Survival Prediction in Primary Glioblastoma. Mol Neurobiol 57:5235–5246. https://doi.org/10.1007/s12035-020-02088-w
68. Xiong Y, Lei J, Zhao J, Feng Y, Qiao T, Zhou Y, et al. Gene expression-based clinical predictions in lung adenocarcinoma. Aging 2020;12:15492–503. https://doi.org/10.18632/aging.103721
69. Bao H, Ren P, Yi L, Lv Z, Ding W, Li C, Li S, Li Z, Yang X, Liang X, Liang P (2023) New insights into glioma frequency maps: From genetic and transcriptomic correlate to survival prediction. Int J Cancer 152(5):998–1012
70. Yang H, Qiu W, Liu Z (2024) Anoikis-related mRNA-lncRNA and DNA methylation profiles for overall survival prediction in breast cancer patients. Math Biosci Eng 21(1):1590–1609
71. Nassani R, Bokhari Y, Alrfaei BM (2023) Molecular signature to predict quality of life and survival with glioblastoma using Multiview omics model. PLoS ONE 18(11):e0287448
72. Subramanian V, Syeda-Mahmood T, Do MN (2024) Modeling-based joint embedding of histology and genomics using canonical correlation analysis for breast cancer survival prediction. Artif Intell Med 1(149):102787
73. Jaksik R, Szumała K, Dinh KN, Śmieja J (2024) Multiomics-Based Feature Extraction and Selection for the Prediction of Lung Cancer Survival. Int J Mol Sci 25(7):3661
74. Mohammed MA, Abdulkareem KH, Dinar AM, Zapirain BG (2023) Rise of Deep Learning Clinical Applications and Challenges in Omics Data: A Systematic Review. Diagnostics 13(4):664
75. Sun D, Wang M, Li A (2019) A Multimodal Deep Neural Network for Human Breast Cancer Prognosis Prediction by Integrating Multi-Dimensional Data. IEEE/ACM Trans Comput Biol Bioinform 16:841–850. https://doi.org/10.1109/TCBB.2018.2806438
76. Chen H, Gao M, Zhang Y, Liang W, Zou X. Attention-Based Multi-NMF Deep Neural Network with Multimodality Data for Breast Cancer Prognosis Model. Biomed Res Int 2019;2019. https://doi.org/10.1155/2019/9523719
77. Tong L, Mitchel J, Chatlin K, Wang MD. Deep learning-based feature-level integration of multi-omics data for breast cancer patients survival analysis. BMC Med Inform Decis Mak 2020;20. https://doi.org/10.1186/s12911-020-01225-8
78. Kirtania R, Banerjee S, Laha S, Shankar BU, Chatterjee R, Mitra S. Deepsgp: Deep learning for gene selection and survival group prediction in glioblastoma. Electronics (Switzerland) 2021;10. https://doi.org/10.3390/electronics10121463
79. Fathi Kazerooni A, Saxena S, Toorens E, Tu D, Bashyam V, Akbari H, et al. Clinical measures, radiomics, and genomics offer synergistic value in AI-based prediction of overall survival in patients with glioblastoma. Sci Rep 2022;12. https://doi.org/10.1038/s41598-022-12699-z
80. Braman N, Gordon JWH, Goossens ET, Willis C, Stumpe MC, Venkataraman J. Deep Orthogonal Fusion: Multimodal Prognostic Biomarker Discovery Integrating Radiology, Pathology, Genomic, and Clinical Data. ArXiv - CS - Multimedia 2021.
81. Sharma D, Deepali, Garg VK, Kashyap D, Goel N. A deep learning-based integrative model for survival time prediction of head and neck squamous cell carcinoma patients. Neural Comput Appl 2022. https://doi.org/10.1007/s00521-022-07615-5
82. Kim Y, Kang JW, Kang J, Kwon EJ, Ha M, Kim YK, et al. Novel deep learning-based survival prediction for oral cancer by analyzing tumor-infiltrating lymphocyte profiles through CIBERSORT. Oncoimmunology 2021;10. https://doi.org/10.1080/2162402X.2021.1904573

83. Tapak L, Ghasemi MK, Afshar S, Mahjub H, Soltanian A, Khotanlou H (2023) Identification of gene profiles related to the development of oral cancer using a deep learning technique. BMC Med Genomics 16(1):35
84. Othman NA, Abdel-Fattah MA, Ali AT (2023) A hybrid deep learning framework with decision-level fusion for breast cancer survival prediction. Big Data and Cognitive Computing 7(1):50
85. Palmal S, Arya N, Saha S, Tripathy S (2023) Breast cancer survival prognosis using the graph convolutional network with Choquet fuzzy integral. Sci Rep 13(1):14757
86. Vollmer A, Hartmann S, Vollmer M, Shavlokhova V, Brands RC, Kübler A, Wollborn J, Hassel F, Couillard-Despres S, Lang G, Saravi B (2024) Multimodal artificial intelligence-based pathogenomics improves survival prediction in oral squamous cell carcinoma. Sci Rep 14(1):5687
87. Jia L, Ren X, Wu W, Zhao J, Qiang Y, Yang Q (2024) DCCAFN: deep convolution cascade attention fusion network based on imaging genomics for prediction survival analysis of lung cancer. Complex & Intelligent Systems 10(1):1115–1130
88. Mvududu NH, Sink CA (2013) Factor Analysis in Counseling Research and Practice. Counseling Outcome Research and Evaluation 4:75–98. https://doi.org/10.1177/2150137813494766
89. Partridge M, Calvo R (1997) Fast dimensionality reduction and simple PCA. Intell Data Anal 2(3):203–14
90. Zou H, Hastie T, Tibshirani R (2006) Sparse principal component analysis. J Comput Graph Stat 15:265–286. https://doi.org/10.1198/106186006X113430
91. Rosipal R, Girolami M, Trejo LJ, Cichocki A (2001) Kernel PCA for feature extraction and de-noising in nonlinear regression. Neural Comput Appl 10:231–243. https://doi.org/10.1007/s521-001-8051-z
92. Muthukrishnan R, Rohini R. LASSO: A feature selection technique in predictive modeling for machine learning. 2016 IEEE International Conference on Advances in Computer Applications, ICACA 2016, Institute of Electrical and Electronics Engineers Inc.; 2017, p. 18–20. https://doi.org/10.1109/ICACA.2016.7887916
93. Han K, Wang Y, Zhang C, Li C, Xu C. Autoencoder Inspired Unsupervised Feature Selection. ICASSP, IEEE International Conference on Acoustics, Speech, and Signal Processing - Proceedings, vol. 2018- April, Institute of Electrical and Electronics Engineers Inc.; 2018, p. 2941–5. https://doi.org/10.1109/ICASSP.2018.8462261
94. Dang T (2019) FSelector: Variable Selection Using Visual Features. In: Graphics Interface, pp 1–9
95. Peng H, Long F, Ding C (2005) Feature selection based on mutual information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. IEEE Trans Pattern Anal Mach Intell 27:1226–1238. https://doi.org/10.1109/TPAMI.2005.159
96. Yu N, Wu MJ, Liu JX, Zheng CH, Xu Y (2021) Correntropy-Based Hypergraph Regularized NMF for Clustering and Feature Selection on Multi-Cancer Integrated Data. IEEE Trans Cybern 51:3952–3963. https://doi.org/10.1109/TCYB.2020.3000799
97. Dormuth I, Liu T, Xu J, Pauly M, Ditzhaus M (2022) A comparative study to alternatives to the log-rank test
98. Wald R, Khoshgoftaar TM, Sloan JC. Using feature selection to determine optimal depth for wavelet packet decomposition of vibration signals for ocean system reliability. Proceedings of IEEE International Symposium on High Assurance Systems Engineering, 2011, p. 236–43. https://doi.org/10.1109/HASE.2011.60
99. Eid HF, Hassanien AE, Kim T hoon, Banerjee S. Linear Correlation-Based Feature Selection for Network Intrusion Detection Model. Communications in Computer and Information Science, vol. 381 CCIS, Springer Verlag; 2013, p. 240–8. https://doi.org/10.1007/978-3-642-40597-6_21
100. Uguroglu S, Carbonell J (2011) Feature Selection for Transfer Learning. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, Berlin, Heidelberg, pp 430–442
101. Hasan H, Tahir NM, Feature selection of breast cancer based on Principal Component Analysis. Proceedings - CSPA (2010) 2010 6th International Colloquium on Signal Processing and Its Applications. IEEE Computer Society 2010:242–245. https://doi.org/10.1109/CSPA.2010.5545298

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.