International Conference on Multidisciplinary Perspectives in Advanced Computing and Technology (IMPACT 2026)
G. B. Pant University of Agriculture and Technology, Uttarakhand, India. Jan. 10-11, 2026

Cancer Disease Prediction using Machine Learning

Mehak Singh, Shambhavi, Amita Sharma

Department of Computer Science and Engineering, Sharda University, Greater Noida, India
singhmehak0210@gmail.com, shambhavim29@gmail.com, safalta.amita@gmail.com

Abstract
Cancer is still among the leading causes of death on the planet. Poor outcomes and higher mortality rates create an urgent need for early diagnosis using high-accuracy cancer-detecting methods. Some diagnostic methods in use include biopsy, imaging, and clinical evaluation. These techniques are highly expensive, time-consuming, and prone to errors due to human involvement. With the advancement of artificial intelligence, machine learning has become one of the strongest tools for predicting, classifying, and diagnosing cancer. Machine learning algorithms will therefore analyze a large volume of patient data, including genetic information, medical images, and patient clinical history, for any unusual patterns that may not be easily spotted by experts. Useful algorithms include Decision Trees, Random Forest, Support Vector Machines, K-Nearest Neighbours, and Deep Learning. This study will focus on selecting appropriate algorithms, preprocessing, and feature extraction to achieve high prediction accuracy. Challenges also persist in data imbalance, overfitting, and ensuring patient data confidentiality. However, the integration of machine learning into cancer diagnosis has been promising, enabling higher early detection rates, supporting personalised treatment planning, and promoting low-cost, efficient, and patient-centred health care solutions.

Keywords: *cancer prediction, machine learning, data preprocessing, early diagnosis, health care AI.*

1. Introduction

Cancer is a multifactorial disorder characterised by the abnormal growth of aberrant cells, which possess the ability to invade all parts of the body and destroy tissues and organs. Cancer is one of the most perilous health hazards in the world, which kills millions of people every year. According to the WHO, cancer is the second most important cause of mortality in the world. The three most common types of cancers are cancers of the breast, lungs, and colorectum. Cancer should be identified as early as possible because it substantially increases the chances of effective treatment and survival.

Conventional techniques for the detection of cancer in the form of imaging tests, biopsies, and laboratory tests are beneficial but possess a number of demerits: they are invasive, time-consuming, and expensive, and may fail to detect cancer at an early stage. Over the last few years, there has been phenomenal development in AI, which has opened new prospects in the medical sector, particularly in disease prediction and diagnosis. ML, a form of AI, has also proved extremely useful in analysing intricate medical data and identifying patterns that can help predict cancer in its early stages. ML algorithms can run on large volumes of patient medical history, genetic information, histopathological images, and clinical diagnostic test reports. The models improve over time and yield accurate, consistent predictions. Applications of ML in cancer research have been successful for predicting cancer development, detecting cancer types, and assisting in treatment planning. Therefore, machine learning is a valuable resource for developing diagnostic systems, helping medical professionals make sound clinical decisions [11-12].

2. Literature Review

Table 1: Comparison of considered papers

S. No.	Year	Authors	Dataset	No. of Records	Algorithm(s)	Result / Accuracy
1.	2019	Lof et al.	Retrospective clinical information from Netherlands Cancer Institute & Amsterdam UMC (serum HE4, CA125, age, WHO performance status)	273 patients (advanced epithelial ovarian cancer)	Multivariate model-derived Cancer Ovarii Non-invasive Assessment of Treatment Strategy (CONATS) index: HE4, CA125, age, WHO score	AUC = 0.80 for prediction of >1 cm residual disease; superior to HE4 (0.76), CA125 (0.72), and age (0.58). In ≥70-year-old patients, AUC = 0.82.
2.	2021	Chen et al.	MJ Health Management Institution, Taiwan (self-paying medical screening cohort, with Cancer Registry and Death File linkage)	234,044 participants (1972 CRC cases, mean follow-up 7.4 years)	Step-wise Cox proportional hazards regression; multiple models combining questionnaire, FIT, and blood biomarkers	Optimal model (questionnaire + blood + FIT) had C-statistic 0.81; accuracy of single FIT model lower and very age-dependent; developed a useful scorecard for CRC risk prediction.
3.	2025	Yahata et al.	Clinical dataset of 164 women (170 nodules) from Southern Brazil (Core Biopsy)	170 records	Multilayer Perceptron (MLP) Artificial Neural Network with SHAP, LIME, PDP, ICE for explainability	ANN had an accuracy of 98.0%; BI-RADS® 5 (ultrasound & mammography), nodule size >2 cm, family history, and age >50 were significant predictors
4.	2025	Pinto et al.	Not applicable (Systematic Review of studies from 2018–2025)	Not Applicable	Review of AI/ML approaches (CNN, EfficientNet, Radiomics, Transformer, cfDNA models, Random Forest, etc.)	Found AI methods have high accuracy (up to 99%), sensitivity (up to 99%), and AUROC (0.962)
5.	2019	Vazifehdan et al.	Omid Hospital Breast Cancer dataset (Iran, 2000–2010); UCI Wisconsin (569 records);	Omid: 217 patients (22 variables, 96 complete records);	Hybrid imputation: Bayesian Network (categorical) + Tensor Factorization (continuous);	For Omid dataset: Accuracy 89.29%, Sensitivity 78.55%, Specificity 92.83% with C4.5; superior quality predictions compared to six other

			UCI Cleveland (303 records)	Wisconsin: 569; Cleveland: 303	classifiers: Decision Tree (C4.5), KNN, SVM	alternative imputation approaches.
6.	2019	Palsdottir et al.	STHLM3-MRI multicentre diagnostic study (Sweden & Norway)	532 men (45–75 years)	Logistic regression models: (1) Stockholm3, (2) modPI-RADS, (3) Unified S3M-MRI (Stockholm3 + modPI-RADS)	Logistic regression models: (1) Stockholm3, (2) modPI-RADS, (3) Unified S3M-MRI (Stockholm3 + modPI-RADS)
7.	2025	Huang et al.	MRI (T2W-MRI) scans + clinical and pathological data from CRC patients (Fujian Medical University Hospital, China)	136 patients	Radiomics feature extraction + ML models (KNN, RF, SVM, LR, XGBoost, LightGBM, MLP); KNN performed best	Clinical radiomics model outperformed single models. Validation AUCs: p53 = 0.758, Syn = 0.739, HER2 = 0.756, PNI = 0.835, VI = 0.797
8.	2023	Chen et al.	MRI with multiple sequences (T1WI, DCE, T2-FS, DWI), First Affiliated Hospital of Soochow University	90 patients (60 train, 30 test) + 20 external validation	Radiomics (189 features: texture, morphology, GLCM) + Naïve Bayes classifier.	Best (combined) model: Train AUC 0.871; Test AUC 0.854. External validation AUC 0.833, Accuracy 0.742, Sensitivity 0.902, Specificity 0.854.
9.	2023	KalaiSelvi et al.	(1)sEVA extracellular-vesicle data with KRAS/P53 mutants; (2) Pan-cancer dataset referenced/used for performance table	1,018 images (after pre processing)	Hybrid model combining CNN (feature extraction) + RNN (temporal/spatial feature learning)	Achieved accuracy= 96.7%, precision = 95.4%, recall = 96.2%, F1-score = 95.8%

10.	2023	Bao et al.	Global Burden of Disease (GBD 2021) database (204 countries, 1990–2021)	Not applicable (aggregated population-level records)	Bayesian Age-Period-Cohort model, decomposition analysis, and inequality indices are: Slope Index and Concentration Index.	Found 216,768 new cases and 171,961 deaths in 2021; incidence increased 101% since 1990; projected ASIR in 2035 = 2.39/100k, ASMR = 1.75/100k

[1] Lof et al. (2019) The fact that the CONATS index was also useful in predicting residual disease following cytoreduction at interval in patients with ovarian cancer was further confirmed by Lof et al., 2019. A model based on HE4 and CA125, age, and performance status maintains its strength at an AUC of 0.80 and 0.82 in elderly patients, respectively, compared to any single biomarker as a non-invasive predictor for surgical planning.

[2] Chen et al. (2021) They have developed a model of colorectal cancer in 234,044 Taiwanese study participants using FIT, questionnaires, and blood tests. In the two-stage model, the C-statistic was 0.81 higher compared to FIT alone. A simple CRC scorecard has been developed for facilitating colonoscopy referral and long-term screening strategy.

[3] Yahata, et al. (2025) The 2025 work of Erika Yahata, Pablo Deoclecia dos Santos, Maria Marlene de Souza Pires, Ricardo Suyama, and Priscyla Waleska Simoes is entitled "Neural Networks and Explainable Artificial Intelligence for Breast Cancer Prediction and Classification." It was based on a clinical dataset consisting of 164 females (170 nodules). A Multilayer Perceptron ANN with SHAP and LIME produced 98% accuracy, the most important predictors being BI-RADS® scores, nodule size, age, and family history.

[4] Pinto et al. (2025) They presented a review of improving health in Africa with a few state-of-the-art ways of predicting and detecting lung cancer using AI-based solutions that include, among others, CNNs, Efficient Net, and Transformers-all with high accuracy as high as 99%. However, this has a number of limitations which includes bias in the datasets, being computationally heavy, and sparse evidence for clinical uptake particularly in the African healthcare systems.

[5] Vazifehdan et al. (2019) They designed a hybrid Bayesian Network+Tensor Factorization to perform missing value imputation with the aim of improving breast cancer recurrence prediction. The proposed method was experimented on with the Omid, Wisconsin, and Cleveland databases. It is said that it achieved 89.29% precision with the C4.5 method, higher than default imputation, and remarkably improved sensitivity, specificity, and overall prediction credibility.

[6] Palsdottir, et al. (2019) This study combined Stockholm3 testing with MRI-based PI-RADS scoring in 532 men to develop the S3M-MRI model. Most accurate combined model compared to single models; AUC 0.88. Clinically useful but with growing advantage over Stockholm3 and MRI sequential use, it is difficult in terms of cost and workflow.

[7] Qiao-yi Huang et al. (2025) Qiao-yi Huang et al., within the research article of the year 2025 "Preoperative prediction of multiple biological characteristics in colorectal cancer using MRI and machine learning," have validated MRI and clinical annotations for 136 patients with CRC. They picked out the radiomics features and reached robust performances with KNN-based integrated models with corresponding AUC values as high as 0.835 when making predictions for PNI.

[8] Chen et al. (2023) in their work, entitled "Prediction of recurrence risk of cervical cancer after radiotherapy using multi-sequence MRI radiomics," tested 90 patients and externally validated 20 cases. Using the Naïve Bayes classifiers, the best overall multi-sequence model was more accurate at 0.854, defeating the single-sequence recurrence prediction models.

[9] B. KalaiSelvi et al. (2023) PSVAM on CNN and IoT was presented by the research study of B. KalaiSelvi et al. published in 2025. This method uses the Pan-Cancer Genomics dataset for detecting multi-cancer. The methodology reached an accuracy of 91.2%, and outperformed SL2MF, HA-GK, IDriveGenes with higher sensitivity and specificity, fast detection, highly successful in early clinical cancer detection.

[10] Bao, et al. (2023) This paper represents a study by Bao et al. (2025) using GBD 2021 data from 204 countries to review trends in gallbladder and biliary tract cancer. Despite worldwide increases in absolute numbers, the standardized mortality and incidence rate declined. Future declines are expected for the year 2035. Sociodemographic patterns, aging, and medical inequity were the major drivers of the global disease burden and disparity.

3. Discussion

Publicly documented and used datasets in the reviewed literature for undertaking predictions of cancer survival, progression, and incidence are mostly from the Breast Cancer Wisconsin dataset, the SEER Dataset, and other hospital or clinic patient records. They generally possess information with respect to demographic measures, clinical measures, tumor measures, biopsy, and imaging procedures. Other than these, respective studies have provided gene expression and genomic data that yield cancer predictions at the molecular level. Their sizes vary from a few hundred to some thousands of patient records. Most of the time, data gets preprocessed in order to handle missing values, scale feature magnitudes, and resample classes to help the model train better. Artificial and surrogate data also feature as databases in research, ostensibly to strengthen models, more so for atypical forms of cancers. Generally, information databases utilized were heterogeneous forms of different cancers whereby mostly time, breast, lung, and colorectal cancers were used most to analyze diversified backgrounds for predictive models of machine learning.

In the literature reviewed, the predictability capability of the used machine and deep learning algorithms was within 75-99%, considering the dataset, kind of cancer, and algorithm used. It is observable that on typical clinical datasets, ensemble methods such as Random Forests and XGBoost are more accurate and consistent but even Logistic Regression and SVM perform well on small sets of data, whereas deep learning models like ANN, CNN, and RNN perform best on high-dimensional imaging and genomic datasets. Overall, this model had superior results in precision, recall, F1-score, and AUC-ROC compared with other models. Some articles underline the great importance of feature preprocessing and feature selection for model performance improvement. The best prediction reliability could typically be noticed in those models that were either trained from gene expression data or from the union of multimodal clinico-radiomics attributes. Some of the studies showed an AUC-ROC >0.95 - a mark of superior malign vs. benign discrimination. Generally, all results provide evidence that well-chosen algorithms may provide highly accurate and clinically useful predictions for prognosis and diagnosis in cancer if enabled by proper preprocessing and feature engineering.

Different machine and deep learning models were applied for the prediction of the occurrence, development, and outcome of cancer. Most of the works have been done using baseline supervised machine learning models in the form of Logistic Regression, Decision Trees, Random Forests, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Naive Bayes, because they can be explained well and do well on structured clinical data. Some ensemble approaches, such as Gradient Boosting, XGBoost, and AdaBoost, were also trending for achieving maximum predictive capability through combination. Deep learning architectures such as ANNs, CNNs, and RNNs have also appeared frequently in articles, mostly in areas that involve either imaging or high-dimensional genomic data. Feature selection and dimensionality reduction techniques such as PCA, RFE, and Information Gain are applied very frequently to enhance model performance and avoid overfitting. Most research studies compared different algorithms on a single database with the aim of pointing out the best and most robust model, along with parameters like accuracy, precision, recall, F1-score, and AUC-ROC. In general, algorithms varied from typical explainable models to complex deep learning architectures.

4. Conclusion and Future Works

Cancer is one of the most destructive health problems in the world, while its early prediction is of prime importance for improving prognosis and mortality rates amongst patients. The paper illustrates the effectiveness of machine learning algorithms in predicting and diagnosing cancers with the help of medical datasets, image datasets, and clinical datasets such as Wisconsin Breast Cancer Dataset (WBCD). These include Random Forests, Support Vector Machines, Artificial Neural Networks, and Deep Learning methods, which have proven very accurate and precise in predicting cancer time and again, with percentages exceeding 95%. The outputs from this study show that ML-based patterns hold great potential to improve diagnostic precision and reduce errors substantially, thereby helping clinicians in decision support systems. The application of ML algorithms in the medical system leads to lower costs, greater accuracy, and timely cancer detection, enabling personalised therapy and better patient care.

Although quite promising, several issues with machine learning's achievements in cancer prediction should be addressed and improved in the future. In the near future, larger and more heterogeneous datasets may be used, integrating multi-centre, real-world data from heterogeneous patient populations, with the intent of increasing model generalizability. A multimodal data approach is necessary to develop more complete predictive models by integrating clinical, genomic, imaging, and lifestyle data. Furthermore, XAI is increasingly crucial for developing interpretable, transparent machine learning models that clinicians can trust and understand, thereby facilitating wider healthcare adoption. Data security and privacy are essential for ethical, privacy-respecting practices, such as federated learning, when handling sensitive medical data. Finally, further research is needed to develop lightweight, scalable ML frameworks that can be easily deployed in real-world, real-time clinical applications, enabling effective cancer screening and diagnosis at health centres. Hybrid and ensemble deep learning models can provide consistent, precise predictive performance by combining multiple algorithms.

References

[1] Lof, P., van den Bult, E., van der Wurff, A., Bakker, H., de Vries, M., Willemse-Koster, R., van der Bos, J. & van der Vlugt, M. (2019) 'Pre-operative prediction of residual disease after interval cytoreduction for epithelial ovarian cancer using HE4', International Journal of Gynecological Cancer, 29(8), pp. 1304-1310. doi:10.1136/ijgc-2019-000581.

[2] Chen, C. H., Tsai, M. K., Wen, C., Wen, C. P. (2021) 'A user-friendly objective prediction model in predicting colorectal cancer based on 234 044 Asian adults in a prospective cohort', ESMO Open, 6(6), p. 100288. doi: 10.1016/j.esmoop.2021.100288.

[3] Yahata, T., Shimizu, H., Takano, A., Suzuki, M., Nakayama, Y. (2025) 'Neural Networks and explainable artificial intelligence for breast cancer prediction and classification', Procedia Computer Science, 256, pp. 1159-1166. doi: 10.1016/j.procs.2025.02.224.

[4] Yousef, R., Gupta, G., Yousef, N., & Khari, M. (2022). ‘A holistic overview of deep learning approach in medical imaging’, Multimedia Systems, 28(3), 881-914.

[5] Vazifehdan, M., Moattar, M.H. and Jalali, M. (2019) ‘A hybrid bayesian network and tensor factorization approach for missing value imputation to improve breast cancer recurrence prediction’, Journal of King Saud University - Computer and Information Sciences, 31(2), pp. 175–184. doi:10.1016/j.jksuci.2018.01.002.

[6] Palsdottir, T., Nordström, T., Aly, M., Jäderling, F., Clements, M., Grönberg, H. & Eklund, M. (2019) 'A unified prostate cancer risk prediction model combining the Stockholm3 test and Magnetic Resonance Imaging', European Urology Oncology, 2(5), pp. 490–496. doi:10.1016/j.euo.2018.09.008.

[7] Sagili, S. R., Chidambaranathan, S., Nallametti, N., Bodele, H. M., Raja, L., & Gayathri, P. G. (2024, July). NeuroPCA: Enhancing Alzheimer’s disorder Disease Detection through Optimized Feature Reduction and Machine Learning. In 2024 Third International Conference on Electrical, Electronics, Information and Communication Technologies (ICEEICT) (pp. 1-9). IEEE.

[8] Chen, L., Zhao, Q., Wang, H., Hu, Z. (2025) 'Prediction of recurrence risk of cervical cancer after radiotherapy using multi-sequence MRI radiomics', Radiation Medicine and Protection, 6(3), pp. 169–174. doi:10.1016/j.radmp.2025.04.001.

[9] Bandla, S. L. (2025). Modeling and Optimization of Blood Circulation for Improved Cardiovascular Health. Authorea Preprints.

[10] Bao, Y., Li, J., Zhang, H., Chen, Y., Wang, S. (2025) 'Trends and cross-country inequalities in the global, regional, and national burden of gallbladder and biliary tract cancer from 1990 to 2021, along with the predictions for 2035', Cancer Epidemiology, 96, p. 102802. doi:10.1016/j.canep.2025.102802.

[11] Bandla, S. L. (2025). Neural Stem Cells and Their Role in Regenerative Therapies for Spinal Cord Injury and Neurodegenerative Diseases. *Authorea Preprints*.

[12] Gaddam, M. K. (2025, September). Architecting Observability for AI-Driven Microservices at Scale. In *2025 3rd International Conference on Intelligent Cyber Physical Systems and Internet of Things (ICoICI)* (pp. 1830-1838). IEEE.