Proceedings of CONF-MLA 2025 Symposium: Intelligent Systems and Automation: AI Models, IoT, and Robotic Algorithms
DOI: 10.54254/2755-2721/2025.LD27262

Machine Learning in Early Cancer Detection: A Review of Methods and Applications

Baoyi Zou

Diamond Bar High School, Diamond Bar, USA
zoubaoyi174@gmail.com

Abstract. Cancer ranks among the diseases with the highest global mortality and complication rates. Early detection of cancer can significantly improve patients' survival rates and burden of treatment costs. However, traditional methods such as Computed Tomography (CT) and Nuclear Magnetic Resonance Imaging (MRI) have a high false positive rate and limited accuracy when detecting cancer in the early stages. In the past few years, machine learning has appeared as a crucial tool for pattern recognition in biomedical data. By combining machine learning with traditional detection methods, new approaches and possibilities for early cancer detection have been explored. This review summarizes and compares the applications of machine learning in the early detection of cancer. The paper discusses the current status and challenges of machine learning in early cancer applications; analyzes the advantages and limitations of common machine learning technology techniques and methods in clinical practice translation; and explores future directions and possible addressing solutions.

Keywords: Machine learning, Early cancer detection, Deep learning, Artificial intelligence.

1. Introduction

Cancer is a major public health issue. Many types of cancer, including lung cancer, breast cancer, and colorectal cancer have high incidence rates. Early detection of cancer in the early stage is a critical factor to improve patient outcomes and mortality rates and reduce treatment costs. In epidemiological study, it demonstrated that survival rates are significantly higher for lung, breast, and colorectal cancers if the disease is diagnosed at an early stage. Traditional screening tools play a significant role for cancer detection. For lung cancer, Low-dose computed tomography (LDCT) is often used. However, it has high false-positive rates which lead patients to do unnecessary checks or surgery. Mammography is a traditional method for detection of breast cancer but lacks sensitivity in patients with dense breast tissue. These methods highly rely on expert interpretations.

Lately, artificial intelligence (AI) has developed rapidly and plays an important role in many areas. Machine learning (ML) excels at identifying high-dimensional and complicated patterns with large medical datasets. The models are able to capture small details and subtle features that humans do not easily notice or observe. Martinez et al. (2023) discovered that deep learning algorithms outperform traditional machine learning methods in the detection of breast cancer in the early stage from sources encompassing X-ray images [1]. The algorithms have demonstrated progress in diverse

DOI: 10.54254/2755-2721/2025.LD27262

data modalities. ML can also recognize important features in data with noise given by liquid biopsy [2]. By using multimodal data integration, AI systems can provide more accurate prediction outcomes and improve sensitivity and specificity. In addition, it might be able to uncover some unnoticed disease signals that are invisible to traditional screening methods. Deep learning models, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs) applied to longitudinal CT scans, can non-invasively track tumour phenotypes [3]. By integrating datasets and models, ML has potential to surpass traditional screening techniques in cancer detection. Despite the advancements, there are some significant challenges remaining in translating AI-based techniques into clinical practice. These issues must be addressed to facilitate the reliable practice of ML-based diagnostics in the cancer research field.

The purpose of the review is comparing performance metrics across different machine learning methods and explaining benefits of combining multiple types of data. Furthermore, it discusses current challenges including clinical practice, limited datasets, and difficulty of model interpretation.

2. Methods

2.1. Data sources

The effectiveness of machine learning largely relies on the quantity and variability of data. Biomedical data sources with a wide range to support the development of predictive models when facing diverse situations. The large datasets provide a strong foundation for the development of applying machine learning in early cancer detection.

Medical imaging is still the most widely used source. The techniques include mammography for breast cancer, computed tomography (CT) for lung cancer, endoscopy for colorectal cancer, etc. These provide visual information so that structural abnormalities can be detected at an early stage. The medical imaging datasets are often used to train CNNs and other deep learning models [4].

Liquid biopsy enables markings such as circulating tumor DNA (ctDNA), circulating tumor cells (CTCs), and microRNAs (miRNA). It provides non-invasive biomarkers and captures molecular alterations [1]. These molecular features can reflect tumor activity in the early stage. The models based on liquid biopsy improve the accuracy of classifying cancerous versus non-cancerous samples.

Genomics and multi-omics datasets also provide valuable insights, such as The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO) [2]. It involves providing different types of data from the same set of samples in order to gain more comprehensive analyses of biological diseases and modeling systems [5].

2.2. Preprocessing and feature engineering

Before applying biomedical data to machine learning, preprocessing and feature engineering are necessary. The purpose of preprocessing and feature engineering is to enhance the quality of data and the performance of models. This step is essential because unprocessed data with noise, imbalance, and heterogeneity.

For medical imaging data, data processing usually includes reducing noise, normalization, and image augmentation. Noise reduction removes irrelevant information and extracts meaningful features [5]. Normalization refers to adjusting pixels into standardized range, making images consistent brightness and contrast. For example, rescaling CT scans from different machines which makes data at a comparable level and reduces bias created by devices. Image augmentation increases

DOI: 10.54254/2755-2721/2025.LD27262

variability of data artificially by rotating, flipping, or scaling images to simulate different scanning conditions.

Dimensionality reduction and feature selection play a crucial role when preprocessing genomic and molecular data. High-dimensional datasets often include thousands of features which many of them are irrelevant or redundant. Principal component analysis (PCA) and least absolute shrinkage, autoencoders and selection operator (LASSO) are three main methods used to identify variables and minimize noise [2].

Another significant challenge is class imbalance. In early cancer detection, positive cases are scarce. The imbalance tends to bias models that predict negative outcomes. To address this problem, Synthetic Minority Oversampling Technique (SMOTE) is widely applied. SMOTE is mainly used for tabular data from genomic or liquid biopsy that generates data points between samples which creates new and reasonable data [2]. Another technique, GANs, generate samples that are similar to realistic data to train between generator and discriminator. GANs involve creating highly complex data such as medical imaging and tissue section [5]. However, SMOTE lacks variability and complexity which tend to lead generative samples to become “average”, loss accuracy rate when identifying some special cases. The training process of GANs is sophisticated which needs massive data and computing resources.

2.3. Machine learning and deep learning protocols

Support Vector Machines (SVM), Gradient Boosting frameworks (XGboost) and Random Forest (RF) are the most common traditional machine learning models [4]. Traditional machine learning remains highly effective when processing non-imaging data. They have been widely applied in liquid biopsy and genomic-based cancer detection [2]. These models can effectively process structured and high-dimensional datasets by providing feature importance. They often serve as baseline models for comparison. Compared to deep learning, traditional machine learning requires less computational resources. However, they underperform when tasks with complex imaging data are given.

For medical imaging, deep learning (DL) is preferred to be used because it automatically chooses exact features. When processing mammography, CT, and endoscopy data, deep learning performs with high sensitivity and specificity. Convolutional Neural Networks (CNNs), including well-known architectures like VGG, ResNet, and DenseNet have been applied to detect tumors and classify malignancies [6]. U-Net, a model designed for locating. It plays a critical role when identifying regions of interest, such as tumors and lesions. It is able to achieve precise localization with manual feature engineering. Based on U-Net, 3D U-Net, Attention U-Net, and Residual U-Net are developed, which are advanced imaging models used for different situations.

Cancer progression is complicated, which is affected by multiple factors. A single model is not able to capture the full picture. Multimodal learning offers a more comprehensive view by integrating diverse types of data.

3. Results

3.1. Evaluation metrics

A set of evaluation metrics is used to determine the performance of models of machine learning for cancer detection [5]. Accuracy measures the proportion of correct prediction that models classify. Specificity evaluates the true negative cases. High specificity means fewer false positive cases.

DOI: 10.54254/2755-2721/2025.LD27262

Sensitivity measures how well when detects actual cancer cases which a high sensitivity imply fewer cancer cases are missed. F1-score measures the balance between precision and sensitivity. A well-developed model can be both accurate and avoid false alarms in detection. Area Under the ROC Curve (AUC) measures how well the models perform in classifying cancerous patients. A model with an AUC closer to 1 indicates it can perfectly separate patients; a model with an AUC of 0.5 indicates identify patients by random guessing.

3.2. Visualization and analysis

For further analysis, a chart was created to compare the functionality and performance of different machine learning and deep learning methods. The visualization used accuracy, specificity, and sensitivity as evaluation metrics to provide a clearer understanding of different methods' strengths and limitations.

The result of Figure 1 shows the advantage of hybrid deep learning frameworks [7-10]. Overall, VGG19+SVM performs the best among these methods. Compared to traditional machine learning models such as SVM, hybrid models have more powerful feature extraction capabilities. SVM requires human manners which make it easy to overlook potential important information. VGG19+SVM, a hybrid model combining the strengths of traditional machine learning and deep learning. VGG19 has a high level of ability for feature extraction; SVM is more stable when dealing with a small sample size and non-linear classification.

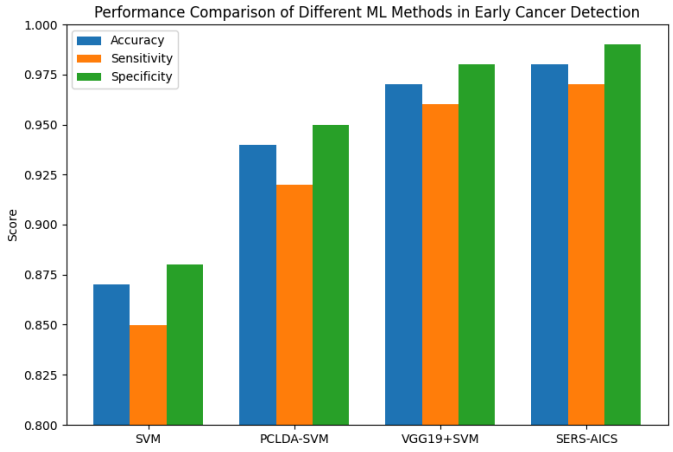

Figure 1. Comparison of the performance of SVM, PCLDA-SVM, VGG19+SVM, and SERS-AICS by evaluating accuracy, sensitivity and specificity

4. Discussion

4.1. Challenges

Data scarcity and imbalance are the most critical issues that need to be addressed. Early-stage cancer cases are rare, which makes it hard to balance negative and positive cases in datasets. Collecting sufficiently large datasets with enough positive samples is difficult. Most cases used to train models are healthy or late-stage cases, which can cause bias and false negative predictions. Even though tools such as GAN and SMOTE solve this problem partially by generating realistic synthetic data, it still affects the reliability of the models. Black-box nature is another concern about ML and DL

DOI: 10.54254/2755-2721/2025.LD27262

models. While deep learning architecture makes predictions with high performance, the decision making process often lacks transparency. Clinicians cannot blindly trust or interpret predictive results without understanding the characteristics that drive decisions. This is a lack of interpretability in medical practice.

4.2. Future directions

Data sharing is essential. Many institutions have valuable datasets but not shared due to patient privacy concerns. If data can be shared while protecting user privacy, the research process will be greatly accelerated. Federated learning might be a solution to collect and integrate data securely. Applying research to clinical practice is crucial which can help clinicians trust and use AI detection model systems effectively. With feedback, models can be further improved. Multimodal integration and large models are key directions that need to be focused on with close attention. Advanced models provide broader perspectives while integrating diverse data types.

5. Conclusion

This review introduced machine learning and deep learning methods in the research fields of early cancer detection. The techniques are powerful tools that reshaped the landscape of cancer detection. They extract patterns and detail from imaging, liquid biopsy, and multi-omics data that are usually overlooked. The article also explores data sources and data preprocessing, which machine learning involves feature engineering. In addition, the paper also analyzes the performance between different methods of machine learning—traditional machine learning and multimodal learning. In comparison, multimodal learning has better performance due to the integrated advantages of multiple methods, which provide a more comprehensive view. However, challenges still remain. Limited availability of datasets, lack of models' interpretability, and class imbalance hinder clinical translation. Algorithm innovation and data sharing are crucial to solving these issues. Computing power is expected to support more accurate, reliable, and accessible cancer screening systems. Machine learning has the potential to transform early cancer detection into a more accurate, affordable, and patient-centred process. The improved machine learning models will play a significant role in clinical practice, improving survival rates and providing a better quality of life for human beings in the future.

References

[1] Gonzales Martinez, R., & van Dongen, D.-M. (2023). Deep learning algorithms for the early detection of breast cancer: A comparative study with traditional machine learning. Informatics in Medicine Unlocked, 41, 101317. https: //doi.org/10.1016/j.imu.2023.101317

[2] Liu, L., Chen, X., Petinrin, O. O., Zhang, W., Rahaman, S., Tang, Z.-R., & Wong, K.-C. (2021). Machine Learning Protocols in Early Cancer Detection Based on Liquid Biopsy: A Survey. Life, 11(7), 638. https: //doi.org/10.3390/life11070638

[3] Xu, Y., Hosny, A., Zeleznik, R., Parmar, C., Coroller, T., Franco, I., Mak, R. H., & Aerts, H. J. W. L. (2019). Deep Learning Predicts Lung Cancer Treatment Response from Serial Medical Imaging. Clinical Cancer Research, 25(11), 3266–3275. https: //doi.org/10.1158/1078-0432.ccr-18-2495

[4] Saba, T. (2020). Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges. Journal of Infection and Public Health, 13(9), 1274–1289. https: //doi.org/10.1016/j.jiph.2020.06.033

[5] Singh, G., Anushka Kamalja, Patil, R., Ashutosh Karwa, Tripathi, A., & Chavan, P. (2024). A comprehensive assessment of artificial intelligence applications for cancer diagnosis. Artificial Intelligence Review, 57(7). https: //doi.org/10.1007/s10462-024-10783-6

DOI: 10.54254/2755-2721/2025.LD27262

[6] Ahmad, I., & Fahad Alqurashi. (2024). Early Cancer Detection Using Deep Learning and Medical Imaging: A Survey. Critical Reviews in Oncology/Hematology, 104528–104528.
[7] Shi, L., Li, Y., & Li, Z. (2023). Early cancer detection by SERS spectroscopy and machine learning. Light: Science & Applications, 12(1), 234. https: //doi.org/10.1038/s41377-023-01271-7
[8] Dubey, P., & Kumar, S. (2023). Advancing prostate cancer detection: a comparative analysis of PCLDA-SVM and PCLDA-KNN classifiers for enhanced diagnostic accuracy. Scientific Reports, 13(1). https: //doi.org/10.1038/s41598-023-40906-y
[9] Mahmoud, N. M., & Soliman, A. M. (2024). Early automated detection system for skin cancer diagnosis using artificial intelligent techniques. Scientific Reports, 14(1), 9749. https: //doi.org/10.1038/s41598-024-59783-0
[10] Dong, S., He, D., Zhang, Q., Huang, C., Hu, Z., Zhang, C., Nie, L., Wang, K., Luo, W., Yu, J., Tian, B., Wu, W., Chen, X., Wang, F., Hu, J., & Xiao, X. (2023). Early cancer detection by serum biomolecular fingerprinting spectroscopy with machine learning. ELight, 3(1). https: //doi.org/10.1186/s43593-023-00051-5