Scholars Journal of Engineering and Technology

OPEN ACCESS

Abbreviated Key Title: Sch J Eng Tech
ISSN 2347-9523 (Print) | ISSN 2321-435X (Online)
Journal homepage: https://saspublishers.com

Studies of Cancer Prediction Using Machine Learning: A Survey

Sonali Mondal Das[1*], Abhoy Chand Mondal[2]

[1]Research Scholar, Department of Computer Science, The University of Burdwan, Burdwan Rajbati, Raiganj, Bardhaman, West Bengal 713104, India
[2]Professor, Department of Computer Science, The University of Burdwan, Burdwan Rajbati, Raiganj, Bardhaman, West Bengal 713104, India

DOI: https://doi.org/10.36347/sjet.2025.v13i02.001 | **Received:** 23.12.2024 | **Accepted:** 29.01.2025 | **Published:** 03.02.2025

***Corresponding author:** Sonali Mondal Das
Research Scholar, Department of Computer Science, The University of Burdwan, Burdwan Rajbati, Raiganj, Bardhaman, West Bengal 713104, India

Abstract **Review Article**

Cancer is a complex diseases that has diversity in terms of its origins, genetic variations and has been categorized as numerous distinct sub types. Early cancer diagnosis and prognosis are essential for patient care and cancer research because they allow for more individualized treatment and better treatment results. So the Machine learning methods invoke model progression and treatment of cancer condition. The machine learning tools are able to extract significant information from intricate data and is indispensable for enhancing decision making and addressing challenges in numerous sectors. Various of techniques, including Bayesian Networks (BNs), Decision Trees, Support Vector Machines (SVMs) and Artificial neural networks (ANNs) are used for the development of predictive model. Therefore, enhancing existing A.I. and M.L. technologies and fostering new advancements are crucial for patient benefit. This literature review delves into the application of A.I. and M.L. algorithms in cancer prognosis, discussing their present use, constraints, and prospects for the future. Clearly our objective is to demonstrate that employing machine learning can advance our understanding of cancer progression, ultimately leading to improved and more accurate decision making in realm of cancer research and treatment.

Keywords: M.L., A.I., Treatment selection, cancer diagnosis, cancer related mortality.

1. INTRODUCTION

Clinicians currently depend on their personal expertise to estimate the survival times of cancer patients, but research has demonstrated that these estimates often prove inaccurate. Particularly, for patients expected to survive longer than three months, only around 60% actually reach this milestone. Studies reveal a tendency to overestimate short-term survival while underestimating long-term survival. Machine learning presents a fitting solution to this challenge due to its capacity to rapidly analyze vast datasets comprising numerous patient cases. Unlike relying solely on clinical expertise, which is limited to individual experiences, machine learning algorithms can process a much broader spectrum of patient data, potentially leading to more precise predictions. As per the GLOBO-CAN DATABASE, lung cancer continues to top the list as the primary cause of cancer-related deaths, with stomach, breast cancer, liver, colorectal, pancreatic cancer, prostate cancer and Lung cancer following closely behind. Cancer ranks second only to heart disease as the leading cause of mortality globally. Artificial intelligence and machine learning together are typically understood to mean a set of algorithms or programs that have been coded into a computer. These technologies provide forecasts and make choices based on data analysis along with embedded instructions.

2. LITERATURE REVIEW

The study explores the use of machine learning (ML) to improve breast cancer diagnosis by analyzing biopsy images. Highlighting the global significance of breast cancer, it evaluates several ML models, including Logistic Regression, Random Forest, Naive Bayes, and Support Vector Machine. By optimizing parameters and combining models using a weighted voting system, the researchers achieve high performance, with 98% accuracy, 97% precision, and 99% recall.

The model, currently a prototype, shows promise for future refinement and application in clinical settings with larger datasets.

This study aims to improve the diagnosis of prostate cancer (PC) by addressing the issues with current biopsy methods, which can sometimes lead to

problems or incorrect results. The researchers use dynamic contrast-enhanced ultrasound (DCE-US) with special microbubble agents to better see blood flow in the prostate, making it easier to find cancerous areas. They develop a new three-dimensional (3D) version of existing imaging techniques that helps predict biopsy results more accurately. By using machine learning models, like Support Vector Machine (SVM) and Gaussian Mixture Model (GMM), they improve the ability to tell the difference between cancerous and non-cancerous tissue. The study also finds that leaving out low-risk cancer cases leads to better diagnosis, pointing out some flaws in traditional methods. Future work will test these findings further through studies involving surgery and targeted biopsies using the new 3D imaging, with the goal of making prostate cancer detection more accurate and effective.

The study implemented various machine learning models, including XG-Boost, achieving 98% accuracy in classifying tumors as malignant or benign based on patient and tumor data. Study includes related studies section. Comparative research on machine learning models conduction. Machine learning models trained on genetic, clinical, and histological data can effectively predict lung cancer probability, aiding in early diagnosis and potentially improving survival rates. Lists machine learning models for lung cancer probability prediction.

Tan *et al*., [7] examined the viability of employing decision stumps, a basic classification method, along with track element analysis, to predict early-stage lung cancer by combining them with Adaboost. In comparing the results, Adaboost outperformed the Fisher Biased Analytic (FDA). This suggests that the combination of Adaboost with urine analysis holds promise as a valuable method for diagnosing early-stage lung cancer in clinical settings.

Kim *et al*., [2] developed a decision tree to analyze occupational lung cancer using data from 1992 to 2007, involving 153 cases reported by the Occupational Safety and Health Research Institute (OSHRI). They examined factors such as smoking history, age, histology, latency period, industry size, sex, exposure and working hours. Utilizing the Classification and Regression Tree (CART) model, consultation with lung disease specialists emerged as the most reliable indicator. However, the CART model's accuracy in determining lung cancer functionality necessitates careful assessment.

Maciej Zieba *et al*., [8] presented enhanced SVM in 2014, which tackles unbalanced data by integrating cost-sensitive support with ensemble classifiers vectors. They used the improved SVM to estimate postsurgery life expectancy in patients with lung cancer and evaluated its efficacy by contrasting its performance with alternative algorithms using unbalanced data.

Petousis *et al*., [5] developed dynamic Bayesian Networks (DBNs) for lung cancer screening using longitudinal data from the NLST LDCT arm. They created five DBNs, incorporating factors like demographics, smoking history, and LDCT screening data. These models accurately predicted individual cancer status over time and outperformed traditional methods like logistic regression and naïve Bayes in discriminating between cancer and non-cancer cases, with an average AUC above 0.75.

Lynch *et al*., [7] has segmented survival of lung cancer patients using a variety of decision aids such as decision trees, support vector machines (SVM), gradient boosting machines (GBM), linear regression, and a custom mix set with the SEER database. Important parameters for comparison included age, gender, tumor size, and tumor level.

In [3] research studies, Deep Convolutional Neural Networks (CNNs) are employed for analyzing medical images and assigning labels. For instance, a study conducted in 2015 utilized a multiscale two-layer CNN to diagnose lung cancer, achieving an accuracy of 86.84%. This research extensively explored three important factors—CNN [4] architecture, dataset characteristics, and transfer learning methods—which had not been thoroughly investigated before.

In citation [6], the study highlights the effectiveness of Support Vector Machines (SVM) when combined with Artificial Neural Networks and Decision Trees, achieving a high precision prediction rate of 92.85%. The research also explores survival analysis in prostate cancer, utilizing ANN, logistic regression and decision trees. Furthermore, the study compares patient data related to colon cancer survival prediction, with neural networks demonstrating greater accuracy.

In research [1], predictive models for breast cancer survival were constructed by employing computational regression with two key data mining methods: ANN and Decision Trees. Evaluation through ten-fold cross-validation showed that Decision Trees (C5) yielded the highest accuracy at 93.6% for the holdout study, followed by logistic regression at 89.2% and ANNs at 91.2%. A promising method for cancer prognostication and detection is provided by the combination of diverse and multidimensional data. This integration shows how several analytical and classification techniques can be applied to produce more precise disease forecasts.

3. FUTURE DIRECTION

This research work has the following objectives:-

1. Analyzing cutting-edge lung cancer prediction methodologies and crafting an innovative model for feature extraction.
2. The symptoms of cancer are examined for early prediction.
3. Advanced Deep learning models are designed and developed for lung cancer prediction.
4. Comparing the suggested model to the traditional model to validate it.

4. DISCUSSION

Cancer prediction through machine learning represents a trans-formative approach in healthcare, revolutionizing early detection, prognosis, and treatment strategies. By leveraging vast datasets encompassing patient demographics, genetic profiles, and medical imaging results, machine learning algorithms can discern intricate patterns indicative of cancer development. The size of data is pivotal with larger dataset having more specific representation of cancer heterogenety accross the population. Through meticulous prepossessing and feature selection, these algorithms identify key factors contributing to cancer risk, enabling the construction of predictive models. These models undergo rigorous validation processes, ensuring robust performance metrics and generalizability across diverse patient populations. Upon validation, integration into clinical practice facilitates timely interventions, personalized treatment plans, and improved patient outcomes. However, challenges persist, including data biases, model interpret ability, and ethical considerations surrounding patient privacy. Addressing these challenges while advancing technologies like genomics and imaging holds the key to unlocking the full potential of ML in cancer prediction, ultimately reshaping the landscape of oncological care.

5. CONCLUSION & FUTURE DIRECTION

In summary, the fusion of ML and deep learning techniques has significantly advanced cancer prognosis by capitalizing on trends like complex algorithm utilization and diverse dataset incorporation. These methods facilitate accurate endpoint forecasts, such as survival rates and treatment responses, across various cancer types. Performance hinges on dataset quality, feature selection, and rigorous validation. Continued research is vital for refining algorithms and enhancing predictive models to improve patient outcomes. While artificial neural networks (ANNs) remain a prevalent method in cancer prediction, several alternative approaches exist for forecasting different cancer types. Beyond ANNs, these approaches encompass diverse machine learning techniques that continue to evolve in improving predictive accuracy. Notably, meticulous design and validation of predictive models are imperative to ensure their effectiveness. Moreover, emphasis should be placed on the planning and execution of experiments, particularly concerning the quality and quantity of biological datasets. By prioritizing these aspects, researchers can enhance the efficiency and reliability of prognostic models across various cancers. Combining diverse, complex data provides a promising technique for cancer prognostication and diagnosis. The application of several analytical and classification techniques is demonstrated by this integration, which leads to more precise disease forecasts.

REFERENCES

1. Demidova, L., Klyueva, I., Sokolova, Y., Stepanov, N., & Tyart, N. (2017). Intellectual approaches to improvement of the classification decisions quality on the base of the SVM classifier. *Procedia Computer Science*, 103, 222–230.
2. Kim, T. W., Koh, D. H., & Park, C. Y. (2010). Decision tree of occupational lung cancer using classification and regression analysis. *Safety and Health at Work*, *1*(2), 140–148.
3. Krishnaiah, V., Narsimha, G., & Chandra, N. S. (2013). Diagnosis of lung cancer prediction system using data mining classification techniques. *International Journal of Computer Science and Information Technologies*, *4*(1), 39–45.
4. Ojala, T., Pietikäinen, M., & Mäenpää, T. (2002). Multiresolution grayscale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(7), 971–987.
5. Petousis, P., Han, S. X., Aberle, D., & Bui, A. A. T. (2016). Prediction of lung cancer incidence on the low-dose computed tomography arm of the national lung screening trial: A dynamic Bayesian network. *Artificial Intelligence in Medicine*, 72, 42–55.
6. Picco, N., Gatenby, R. A., & Anderson, A. R. A. (2016). Stem cell plasticity and niche dynamics in cancer progression. *IEEE Transactions on Biomedical Engineering*, *64*(3), 528–537.
7. Tan, C., Chen, H., & Xia, C. (2009). Early prediction of lung cancer based on the combination of trace element analysis in urine and an AdaBoost algorithm. *Journal of Pharmaceutical and Biomedical Analysis*, *49*(3), 746–752.
8. Banerjee, S., & Mondal, A. C. (2024). An intelligent approach towards plant leaf disease detection through different convolutional neural networks. *International Journal of Intelligent Systems and Applications in Engineering*, *12*(2), 536–546.
9. Banerjee, S., & Mondal, A. C. (2024). A sophisticated approach to soil productivity detection using a convolutional neural network-based model. *International Journal of Advanced and Applied Sciences*, *11*(8), 198–210.
10. Banerjee, S., & Mandal, A. C. (2024). An innovative approach involves machine learning algorithms to forecast future farmer revenue. *SSRG International Journal of Electrical and Electronics Engineering*,

11(8), 59–71. https://doi.org/10.14445/23488379/IJEEE-V11I8P106

11. Banerjee, S., Das, S., & Mondal, A. C. (2024). Classification of healthy and diseased broccoli leaves using a custom deep learning CNN model. *International Journal of Innovative Research in Computer Science and Technology (IJIRCST)*, *12*(5), 110–116. https://doi.org/10.55524/ijircst.2024.12.5.15
12. Banerjee, S., Palsani, D., & Mondal, A. C. (2024). Nutritional content detection using vision transformers: An intelligent approach. *International Journal of Innovative Research in Engineering and Management (IJIREM)*, *11*(6), 21–27. https://doi.org/10.55524/ijirem.2024.11.6.3
13. Banerjee, S., Das, S., & Mondal, A. C. (2024). A study of the application domain of large language models in the agricultural sector. *International Journal of Innovative Research in Computer Science and Technology (IJIRCST)*, *12*(5), 74–78. https://doi.org/10.55524/ijircst.2024.12.5.10
14. Banerjee, S., & Mondal, A. C. (2023). An ingenious method for estimating future crop prices that emphasises machine learning and deep learning models. *International Journal of Information Technology (Singapore)*, 15, 4291–4313. https://doi.org/10.1007/s41870-023-01474-6