International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Pre-existing Statistical Predictive Models for Early Detection of Cancer: A Comprehension review

Ankit Babu[1], Sunil Kumar[2], Alok Kumar Singh[3]

[1]Research Scholar, Department of Statistics
[2]Professor & Head, Department of Statistics
[3]Assistant Professor, Department of Statistics
R. B. S. College, Agra, (U.P.) 282002
[1]ankitvarshney141096@gmail.com

Abstract:

Early detection of cancer significantly enhances treatment success and survival rates. A wide range of statistical and machine learning models have been developed to identify high-risk individuals and detect malignancies at early stages. This review provides a comprehensive overview of established predictive models, including logistic regression, Cox proportional hazards models, Bayesian networks, and machine learning techniques such as random forests, support vector machines, and deep neural networks. We examine each model's mathematical foundation, clinical applications, and predictive performance, highlighting their respective advantages and limitations. Common challenges such as data heterogeneity, interpretability, and model generalizability are also discussed. The review concludes with future directions aimed at improving model integration, transparency, and real-world clinical impact in cancer early detection.

Keywords: Predictive Models, Detection of Cancer, logistic Regression and Machine Learning**.**

1. Introduction

Cancer remains one of the leading causes of morbidity and mortality worldwide, accounting for nearly 10 million deaths annually (Sung et al., 2021). The prognosis of cancer is heavily stage-dependent, with early detection significantly improving 5-year survival rates across multiple cancer types, including breast (90% vs. 27% for late-stage), colorectal (90% vs. 14%), and lung (56% vs. 5%) (American Cancer Society, 2023) Given this stark contrast, there is a critical need for robust, generalizable statistical models that can identify high-risk individuals and detect malignancies at their earliest, most treatable stages.

Traditional cancer screening methods, such as mammography, colonoscopy, and low-dose computed tomography (LDCT), while effective, suffer from limitations including high false-positive rates, invasiveness, and accessibility barriers (Esserman et al., 2014). Statistical and machine learning models have emerged as powerful tools to refine screening protocols by incorporating demographic, clinical, genetic, and imaging data into risk stratification algorithms. These models leverage **regression**

techniques, survival analysis, Bayesian inference, and supervised learningto optimize sensitivity (true positive rate) while minimizing unnecessary interventions.

Reviews

The Role of Statistical Modeling in Early Cancer Detection

1. Logistic Regression and Risk Prediction Models

Logistic regression, a generalized linear model (GLM), is widely used for binary classification in oncology (Hosmer et al., 2013). The Gail Model, one of the earliest and most validated statistical tools, estimates breast cancer risk using covariates such as age, family history, and reproductive factors (Gail et al., 1989). The model computes the log-odds of cancer incidence as:

$$\log(p/1-p) = \beta_0+\beta 1_1 X_1 +\cdots\cdots\cdots\cdots\cdots+\beta_k X_k$$

Where, p is the probability of cancer, β is regression coefficients, and X are predictor variables.

Applications:

- **Gail Model (Breast Cancer)**: Estimates 5-year and lifetime risk using age, family history, and hormonal factors.
- **PCPT Risk Calculator (Prostate Cancer)**: Predicts prostate cancer risk based on PSA levels, age, and race.

Strengths:

- Simple, interpretable coefficients.
- Well-suited for clinical risk assessment.

Limitations:

- Assumes linearity between predictors and log-odds.
- Struggles with high-dimensional data (e.g., genomics).

2. Time-to-Event Analysis: Cox Proportional Hazards Models

For cancers where **temporal progression** is critical (e.g., survival analysis), the Cox proportional hazards (PH) model is the gold standard (Cox, 1972). It estimates the hazard function h(t), representing the instantaneous risk of cancer diagnosis or death:

$$h(t\backslash X) = h_0(t)\exp(\beta 1_1 X_1 +\cdots\cdots\cdots\cdots\cdots+\beta_k X_k)$$

Where, h(t)hazard function at time t, $h_0(t)$ is the baseline hazard, and β coefficients quantify the impact of covariates (e.g., tumor stage, biomarkers).

Applications:

- **Nomograms in Colorectal Cancer**: Predicts recurrence risk post-surgery.
- **Lung Cancer Survival Models**: Estimates prognosis based on tumor stage and biomarkers.

Strengths:

- Handles censored data (common in cancer studies).
- Provides hazard ratios for risk factors.

Limitations:

- Requires proportional hazards assumption.

  Less accurate for non-linear relationships

The Cox model is semi-parametric, requiring no assumption about the baseline hazard distribution. However, it assumes proportional hazards, meaning the effect of predictors is constant over time an assumption often violated in real-world data (Grambsch&Therneau, 1994).

3. Bayesian Approaches for Personalized Risk Assessment

Bayesian networks (BNs) provide a **probabilistic graphical framework**to model cancer risk by incorporating prior knowledge (e.g., genetic mutations) and updating beliefs with new evidence (Korb& Nicholson, 2010). For example, the BRCAPRO model computes the posterior probability of BRCA1/2 mutations given family history (Berry et al., 1997):

$P(BRCA|Data) \propto P(Data|BRCA) \cdot P(BRCA)$

Applications:

- **BRCA Mutation Risk (Ovarian/Breast Cancer)**: Integrates family history and genetic testing.
- **Pancreatic Cancer Early Detection**: Combines clinical and imaging data.

Strengths:

- Incorporates prior knowledge (e.g., expert-defined risk factors).
- Handles missing data well.

Limitations:

- Computationally intensive for large datasets.
- Requires expert input for structure learning.

BNs excel in handling missing data and integrating multi-source inputs (e.g., EHRs, imaging). However, they require expert-defined structures and are computationally intensive for large datasets.

4. Machine Learning: From Random Forests to Deep Learning

With the rise of **high-throughput genomics** and **medical imaging**, machine learning (ML) models have surpassed traditional methods in predictive accuracy. Key approaches include:

- **Random Forests (RF)**: An ensemble method aggregating decision trees via **bootstrap aggregation (bagging)**. RFs handle non-linear interactions and high-dimensional data effectively, achieving AUCs >0.90 in lung nodule classification (Hawkins et al., 2016).

- **Support Vector Machines (SVM)**: Maximizes the **margin**between cancer/non-cancer classes using kernel tricks. SVMs perform well in dermatology (melanoma detection, AUC=0.89; Esteva et al., 2017).
- **Neural Networks (NNs):Deep learning**models, particularly CNNs, dominate image-based cancer detection (e.g., mammography, pathology slides). For instance, Google Health's CNN achieved an AUC of 0.96 in breast cancer screening (McKinney et al., 2020).

Despite their power, ML models face challenges in interpretability (e.g., "black-box" NNs) and over-fitting without sufficient training data.

Challenges and Future Directions

1. Data Heterogeneity and Missingness

Real-world datasets often contain censored observations (e.g., lost-to-follow-up patients) and missing covariates, complicating model training. Imputation techniques (e.g., multiple imputations by chained equations, MICE) are commonly employed but may introduce bias (Sterne et al., 2009).

2. Model Calibration and Generalizability

Many models exhibit excellent discrimination (high AUC) but poor calibration (predicted probabilities ≠ observed rates). The Brier score and Hosmer-Lemeshow test are used to assess calibration (Steyerberg et al., 2010). Additionally, external validation is critical a model trained on European populations may fail in Asian cohorts due to genetic/environmental differences (Justice et al., 2019).

3. Performance Evaluation Metrics

- **AUC-ROC**: Measures discriminative power (ideal = 1.0).
- **Sensitivity/Recall**: Critical for minimizing false negatives.
- **Precision/Specificity**: Reduces false positives in screening.
- **Calibration**: Ensures predicted probabilities match observed rates (e.g., Brier score).

4. Ethical and Regulatory Considerations

- AI/ML models must address algorithmic bias (e.g., underperformance in racial minorities; Obermeyer et al., 2019) and comply with GDPR and FDA guidelines for clinical deployment.

Table 1: Comparative Analysis of Statistical Predictive Models for Early Cancer Detection

Model	Cancer Type	Key Features	Performance (AUC-ROC)	Strengths	Limitations	References
Logistic Regression	Breast (Gail Model)	Age, family history, hormonal factors	0.60 - 0.70	Simple, interpretable	Limited to linear relationships	Gail et al. (1989)
	Prostate (PCPT Risk)	PSA levels, age, race	0.68 - 0.75	Clinically adopted	Misses non-linear interactions	Thompson et al. (2006)
Cox PH Model	Colorectal	Tumor stage, biomarkers	0.75 - 0.82	Handles survival data	Requires proportional hazards assumption	Kattan et al. (2003)
	Lung	Smoking history, genetic mutations	0.72 - 0.80	Time-to-event analysis	Sensitive to censoring bias	Mandrekar et al. (2011)
Bayesian Networks	Ovarian (BRCA)	Family history, genetic testing	0.78 - 0.85	Incorporates expert knowledge	Computationally intensive	Antoniou et al. (2008)
	Pancreatic	Imaging + clinical data	0.70 - 0.80	Handles missing data	Requires structured data inputs	Klein et al. (2013)
Random Forest	Breast (Genomics)	Gene expression, mutations	0.85 - 0.90	Robust to overfitting	Less interpretable	Díaz-Uriarte et al. (2006)
	Lung (CT Nodules)	Radiomic features	0.88 - 0.93	High accuracy	Computationally expensive	Hawkins et al. (2016)
SVM	Melanoma	Dermatoscopic images	0.83 - 0.89	Works well with small datasets	Sensitive to kernel choice	Esteva et al. (2017)
	Brain (MRI)	Tumor texture, shape	0.81 - 0.87	Effective in high dimensions	Poor calibration	Zacharaki et al. (2009)
Neural Networks	Lung (X-rays)	CNN-based detection	0.92 - 0.96	State-of-the-art accuracy	Needs large datasets	Ardila et al. (2019)
	Prostate (Pathology)	Deep learning on biopsies	0.89 - 0.94	Automates feature extraction	Black-box nature	Campanella et al. (2019)

Conclusion

Early detection is critical to improving cancer outcomes, and predictive modeling has become a vital tool in this effort. Traditional statistical models like logistic regression and Cox proportional hazards offer interpretability and clinical relevance, while Bayesian networks add flexibility through probabilistic reasoning. More recently, machine learning approaches such as random forests, support vector machines, and deep neural networks have demonstrated impressive performance, particularly in handling high-dimensional data like genomics and medical imaging.

However, challenges such as data heterogeneity, limited generalizability, and ethical concerns around algorithmic bias persist. Future efforts should focus on integrating the strengths of both statistical and machine learning methods, ensuring robust validation and adherence to transparent reporting standards. Collaboration across disciplines will be key to advancing predictive models from research into routine clinical practice, ultimately enhancing personalized cancer prevention and early diagnosis.

References

1. American Cancer Society. (2023). Cancer facts & figures 2023. https://www.cancer.org/research/cancer-facts-statistics.html
2. Antoniou, A. C., Cunningham, A. P., Peto, J., Evans, D. G., Lalloo, F., Narod, S. A., Risch, H. A., Eyfjord, J. E., Hopper, J. L., Southey, M. C., Olsson, H., Johannsson, O., Borg, Å., Pasini, B., Radice, P., Manoukian, S., Eccles, D. M., Tang, N., Olah, E., Easton, D. F. (2008). The BOADICEA model of genetic susceptibility to breast and ovarian cancers: Updates and extensions. American Journal of Human Genetics, 82(2), 465–476. https://doi.org/10.1016/j.ajhg.2007.10.004
3. Ardila, D., Kiraly, A. P., Bharadwaj, S., Choi, B., Reicher, J. J., Peng, L., Tse, D., Etemadi, M., Ye, W., Corrado, G., Naidich, D. P., &Shetty, S. (2019). End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. Nature Medicine, 25(6), 954–961. https://doi.org/10.1038/s41591-019-0447-x
4. Berry, D. A., Iversen, E. S., Gudbjartsson, D. F., Hsu, L.,Scheuner, M. T., Molina, A.,&Parmigiani, G. (1997). BRCAPRO validation, sensitivity analysis and model enhancements. Breast Cancer Research and Treatment, 45(3), 261–271. https://doi.org/10.1023/A:1005888723282
5. Bi, W. L., Hosny, A., Schabath, M. B., Giger, M. L., Birkbak, N. J., Mehrtash, A., ...&Aerts, H. J. (2019). Artificial intelligence in cancer imaging: Clinical challenges and applications. CA: A Cancer Journal for Clinicians, 69(2), 127–157. https://doi.org/10.3322/caac.21552
6. Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., & Fuchs, T. J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nature Medicine, 25(8), 1301–1309. https://doi.org/10.1056/NEJMoa1905041
7. Char, D. S., Shah, N. H., & Magnus, D. (2018). Implementing machine learning in health care—addressing ethical challenges. New England Journal of Medicine, 378(11), 981–983. https://doi.org/10.1056/NEJMp1714229

8. Chaudhary, K., Poirion, O. B., Lu, L., &Garmire, L. X. (2018). Deep learning-based multi-omics integration robustly predicts survival in liver cancer. Clinical Cancer Research, 24(6), 1248–1259. https://doi.org/10.1158/1078-0432.CCR-17-0853
9. Cristianini, N., &Shawe-Taylor, J. (2020). An introduction to support vector machines and other kernel-based learning methods (2nd ed.). Cambridge University Press.
10. Cox, D. R. (1972). Regression models and life-tables. Journal of the Royal Statistical Society: Series B (Methodological), 34(2), 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x
11. Díaz-Uriarte, R., & Alvarez de Andrés, S. (2006). Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7(1), 3. https://doi.org/10.1186/1471-2105-7-3
12. Esserman, L. J., Thompson, I. M., & Reid, B. (2014). Overdiagnosis and overtreatment in cancer: An opportunity for improvement. JAMA, 310(8), 797–798. https://doi.org/10.1001/jama.2013.108415
13. Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., &Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118. https://doi.org/10.1038/nature21056
14. Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ...& Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z
15. Gail, M. H., Brinton, L. A., Byar, D. P., Corle, D. K., Green, S. B., Schairer, C., &Mulvihill, J. J. (1989). Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. Journal of the National Cancer Institute, 81(24), 1879–1886. https://doi.org/10.1093/jnci/81.24.1879
16. Grambsch, P. M., &Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. Biometrika, 81(3), 515–526. https://doi.org/10.1093/biomet/81.3.515
17. Hawkins, S. H., Korecki, J. N., Balagurunathan, Y., Basu, S., Goldgof, D. B., Hall, L. O., ...& Gillies, R. J. (2016). Predicting malignant nodules from screening CT scans. Journal of Thoracic Oncology, 11(12), 2120–2128. https://doi.org/10.1016/j.jtho.2016.08.124
18. Hosmer, D. W., Lemeshow, S., &Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Wiley. https://doi.org/10.1002/9781118548387
19. Justice, A. C., Covinsky, K. E., & Berlin, J. A. (1999). Assessing the generalizability of prognostic information. Annals of Internal Medicine, 130(6), 515–524. https://doi.org/10.7326/0003-4819-130-6-199903160-00016
20. Kattan, M. W., Hess, K. R., Amin, M. B., Lu, Y., Moons, K. G., Gershenwald, J. E., ... & American Joint Committee on Cancer. (2003). Nomograms for rectal cancer. Journal of Clinical Oncology, 21(21), 3633–3638. https://doi.org/10.1200/JCO.2003.05.073
21. Klein, A. P., Lindström, S., Mendelsohn, J. B., Steplowski, E., Arslan, A. A., Bueno-de-Mesquita, H. B., ...& Kraft, P. (2013). An absolute risk model to identify individuals at elevated risk for pancreatic cancer in the general population. Cancer Research, 73(21), 6369–6377. https://doi.org/10.1158/0008-5472.CAN-12-3820
22. Korb, K. B., & Nicholson, A. E. (2010). Bayesian artificial intelligence (2nd ed.). CRC Press.
23. Liu, X., Rivera, S. C., Moher, D., Calvert, M. J., &Denniston, A. K. (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension. BMJ, 370, m3164. https://doi.org/10.1136/bmj.m3164

24. Liu, Y., Chen, P. H. C., Krause, J., & Peng, L. (2019). How to read articles that use machine learning: Users' guides to the medical literature. JAMA, 322(18), 1806–1816.
25. Liu, Y., Chen, P. H. C., Krause, J., & Peng, L. (2019). How to read articles that use machine learning: Users' guides to the medical literature. JAMA, 322(18), 1806–1816. https://doi.org/10.1001/jama.2019.16489
26. Luo, W., Phung, D., Tran, T., Gupta, S., Rana, S., Karmakar, C., ...&Venkatesh, S. (2016). Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view. Journal of Medical Internet Research, 18(12), e323. https://doi.org/10.2196/jmir.5870
27. Mandrekar, S. J., Schild, S. E., Hillman, S. L., Allen, K. L., Marks, R. S., Mailliard, J. A., ... & Jett, J. R. (2011). A prognostic model for advanced stage nonsmall cell lung cancer. Journal of Thoracic Oncology, 6(6), 1034–1038. https://doi.org/10.1097/JTO.0b013e3182037b3f
28. McCradden, M. D., Joshi, S., Mazwi, M., & Anderson, J. A. (2020). Ethical limitations of algorithmic fairness solutions in health care machine learning. The Lancet Digital Health, 2(5), e221–e223. https://doi.org/10.1016/S2589-7500(20)30065-0
29. McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., ...&Suleyman, M. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6
30. Norgeot, B., Quer, G., Beaulieu-Jones, B. K., Torkamani, A., Dias, R., Gianfrancesco, M., ...& Butte, A. J. (2020). Minimum information about clinical artificial intelligence modeling: The MI-CLAIM checklist. Nature Medicine, 26(9), 1320–1324. https://doi.org/10.1038/s41591-020-1041-y
31. Obermeyer, Z., Powers, B., Vogeli, C., &Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447–452. https://doi.org/10.1126/science.aax2342
32. Parikh, R. B., Teeple, S., &Navathe, A. S. (2019). Addressing bias in artificial intelligence in health care. JAMA, 322(24), 2377–2378. https://doi.org/10.1001/jama.2019.18058
33. Steyerberg, E. W., Vickers, A. J., Cook, N. R., Gerds, T., Gonen, M., Obuchowski, N., ...&Kattan, M. W. (2010). Assessing the performance of prediction models: A framework for traditional and novel measures. Epidemiology, 21(1), 128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2
34. Sterne, J. A. C., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., ...& Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: Potential and pitfalls. BMJ, 338, b2393. https://doi.org/10.1136/bmj.b2393
35. Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: A Cancer Journal for Clinicians, 71(3), 209–249. https://doi.org/10.3322/caac.21660
36. Thompson, I. M., Ankerst, D. P., Chi, C., Goodman, P. J., Tangen, C. M., Lucia, M. S., ...&Coltman, C. A. (2006). Assessing prostate cancer risk: Results from the Prostate Cancer Prevention Trial. Journal of the National Cancer Institute, 98(8), 529–534. https://doi.org/10.1093/jnci/djj305
37. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56. https://doi.org/10.1038/s41591-018-0300-7
38. Vollmer, S., Mateen, B. A., Bohner, G., Király, F. J., Ghani, R., Jonsson, P., ...& Hemingway, H. (2020). Machine learning and artificial intelligence research for patient benefit: 20 critical questions

on transparency, replicability, ethics, and effectiveness. BMJ, 368, l6927. https://doi.org/10.1136/bmj.l6927

39. Way, G. P., Sanchez-Vega, F., La, K., Armenia, J., Chatila, W. K., Luna, A., ...& Schultz, N. (2018). Machine learning detects pan-cancer RAS pathway activation in the Cancer Genome Atlas. Cell Reports, 23(1), 172–180. https://doi.org/10.1016/j.celrep.2018.03.046
40. Wijnberge, M., Geerts, B. F., Hol, L., Lemmers, N., Mulder, M. P., Berge, P., ...&Vlaar, A. P. (2020). Effect of a machine learning-derived early warning system for intraoperative hypotension vs standard care on depth and duration of intraoperative hypotension during elective noncardiac surgery. JAMA, 323(11), 1052–1060. https://doi.org/10.1001/jama.2020.0592
41. Zacharaki, E. I., Wang, S., Chawla, S., SooYoo, D., Wolf, R., Melhem, E. R., &Davatzikos, C. (2009). Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme. IEEE Transactions on Medical Imaging, 28(8), 1266–1277. https://doi.org/10.1109/TMI.2008.2004426