International Journal of Integrated Health Sciences (IIJHS)
ISSN Print: 2302-1381; ISSN Online: 2338-4506

Original Article

AI-powered Somatic Cancer Cell Analysis for Early Detection of Metastasis: The 62 Principal Cancer Types

Sandile Buthelezi,[1] Solly Matshonisa Seeletse,[1] Taurai Hungwe,[2] Vimbai Mbirimi-Hungwe[3]

[1]Department of Statistical Sciences, Pretoria, South Africa
[2]Department of Computer Science and Information Technology, Pretoria, South Africa
[3]Department of Academic Literacy and Science Communication Sefako Makgatho University of Health Sciences, Pretoria, South Africa

Article History

Received: August 05, 2024
Accepted: April 03, 2025
Published: April 30, 2025

DOI: 10.15850/ijihs.v13n1.4061
IJIHS. 2025;13(1):8–16

Correspondence:
Sandile Buthelezi,
Department of Statistical Sciences, Sefako Makgatho University of Health Sciences, Pretoria, South Africa.
E-mail: jsbuthelezi@gmail.com

Abstract

Background: Early detection of metastasis is critical in improving survival outcomes in cancer patients, with artificial intelligence offering advanced tools for predictive analytics.

Objective: To emphasize the importance of early metastasis detection in improving cancer patient outcomes, and to highlight that recent advancements in AI-powered somatic cancer cell analysis may enhance early detection and personalize treatment strategies.

Methods: This study leveraged a comprehensive survival and artificial intelligence (AI) powered analysis to identify key genomic and clinical factors influencing cancer prognosis, with a focus on early metastatic detection. The AI algorithms explored the possibility of detecting tumors with a high spread risk. The study underscored the critical role of AI-powered analysis in the early detection of metastasis and the personalization of treatment strategies in cancer care.

Results: By leveraging advanced AI algorithms, key predictors of cancer prognosis such as fraction genome alteration, primary tumor site, and smoking history, all of which significantly influence metastasis outcomes, were identified. Furthermore, the models demonstrated exceptional predictive accuracy, with XGBoost and Support Vector Machines achieving an accuracy of 0.95.

Conclusion: Integrating AI capabilities into clinical workflows holds the promise of significantly enhancing early detection and treatment of metastatic cancer, thereby improving patient outcomes and optimizing therapeutic interventions.

Keywords: Cancer, early detection, machine learning, metastasis

Introduction

Metastasis is the hallmark of cancer growth and is responsible for the majority of cancer-related fatalities. However, it requires earlier detection and better understanding. The rapid growth of cancer biology research and the rise of new paradigms in the study of metastasis have revealed some of the molecular underpinnings of this spreading process.[1] Cancer emerges from a series of molecular events that fundamentally alter the standard properties of cells. The mutated cells divide and grow in the presence of signals that generally inhibit average cell growth.

The growing mutated cells develop new characteristics, including changes in cell structure, decreased cell adhesion, and production of new enzymes.[2] Metastatic cancer growth occurs when cancer cells break from the primary tumor, spread through the body's circulation or lymph vessels, and form new

This is an Open Access article licensed under the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original author and source are properly cited

tumors. This process can occur in three ways: cells can move through the circulation system to distant areas, grow into the surrounding tissue, or travel through the lymph system to nearby or removed lymph nodes.[3]

Martin et al.[4] emphasized that metastasis remains a leading cause of cancer mortality and is increasingly the focus of scientific and clinical investigations. However, the mechanisms remain inadequately understood and strategies in combatting metastasis stay constrained. Yang et al.[5] discovered a new migration mechanism called collective cell migration in many cancers, which can occur as clusters with the tight cell-cell junction in the tumor microenvironments. This migration has been shown to have higher invasive capacity and resistance to clinical treatments than single tumor cell migration. Collective cell migration has been detected in the early stages of cancer patients, highlighting the importance of early disease screenings.

The classical view of tumor metastasis suggests that tumor cell migration begins with a single cell and progresses through various methods before reaching distant tissues and organs. Zhang et al.[6] found that cancer cell collective invasion is regulated by the energetic states of leader-follower cells. Leader cells require more energy than follower cells, and forward invasion consumes and depletes their available energy. A follower cell then takes over the leader position to sustain invasion. This suggests that metabolic pathways can also be repressed by focusing on metabolic pathways, which is a major clinical interest in treating malignant growth. Even though discoveries have been made, Huang[7] stated that the complexity of the metastatic process has made it difficult to gain a full comprehension of the origins of this most lethal aspect of cancer.

Yu et al.[8] further stressed that early metastasis is often misdiagnosed, contributing to poor prognosis and reduced survival. They highlighted the need for improved detection techniques and predictive models to identify early-onset metastatic disease, emphasizing the role of clinicopathological and molecular profiling in achieving this goal.

This study aims to identify and evaluate the most significant factors contributing to cancer metastasis by integrating classical statistical methods with advanced machine learning algorithms. By analyzing 62 principal cancer types, the objective is to develop a superior predictive model that enhances accuracy and precision in forecasting metastatic outcomes.

Methods

This study utilizes a dataset from a clinical study on 62 metastatic cancer types, derived from Zehir et al.,[9] encompassing genomic alterations in 5,193 female and 5,143 male patients, covering 361 distinct tumor types. The data was collected as part of the MSK-IMPACT initiative at Memorial Sloan Kettering Cancer Center (MSKCC), New York, in June 2017. TMSK-IMPACT is a large-scale, prospective clinical sequencing program that combines clinical, genomic, and pathological information. Data preprocessing and analysis were performed using Python. As part of the data preprocessing pipeline, missing values were handled using mean/median imputation, ensuring a consistent and unbiased representation of the feature set prior to normalization and modeling.

For statistical analysis, the Cox Proportional Hazards (CPH) model was employed to assess survival outcomes across the 62 cancer types. The CPH model assumes proportional hazard functions and a linear relationship between the logarithm of the hazard and the covariates.[10] It was used to identify key prognostic factors and to quantify the impact of various biomarkers on patient survival. Feature Selection (FS) was employed to reducing data dimensionality to enhance Machine Learning (ML) algorithms performance This technique detects attribute dependencies and provides a unified view of attribute estimation in regression and classification.[11] A hybrid method was developed by combining FS with ML. Model training used a 70:30 train-test split, and a confusion matrix was used to evaluate performance.

AI-powered algorithms such as Random Forest (RF), Extreme Gradient Boosting (XGBoost), Neural Networks (NN), and Support Vector Machines (SVMs) are used to analyze complex data patterns, improve predictive accuracy, and classify cancerous cells. RF enhances cancer diagnosis accuracy by constructing ensembles of decision trees trained on various datasets, capturing complex features, and enhancing robustness against outliers and noise.[12] XGBoost enhances predictive accuracy by iteratively boosting decision trees, capturing complex interactions and non-linear relationships, optimizing FS, improving model performance, and identifying early metastasis indicators.[13] NN are used to learn complex patterns in high-dimensional cancer data, improving predictive accuracy for early intervention strategies. The

NN consist of three layers: input, hidden, and output, with performance influenced by their structure.[14] SVMs are advanced predictive modelling algorithms that classify complex data, improve metastasis prediction accuracy, and identify biomarkers in oncology, focusing on Structural Risk Minimization (SRM) and higher-dimensional space mapping.[15]

The confusion matrix (CM) is used to measure method performance in classification. It compares the system's classification results with true results. Four conditions are included: True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). Accuracy, precision, and recall are estimated based on these conditions. Accuracy measures the accuracy of the framework in classifying data, precision reflects the degree of accuracy between data and the framework's response, and recall measures the framework's speed in recovering data.

Accuracy in ML and predictive modelling is the proportion of correctly classified instances, crucial for assessing model performance, such as cancer detection algorithms.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} * 100\%$$

Precision is a metric evaluating the accuracy of a model's positive predictions, calculated as the ratio of true positives to the total number of positive predictions.

$$\text{Precision} = \frac{TP}{TP+FP} * 100\%$$

Recall, or sensitivity, measures the percentage of positive cases correctly identified by a model, with high recall indicating its effectiveness in detecting all relevant instances.

$$\text{Recall} = \frac{TP}{TP+FN} * 100\%$$

The F-measure, or F1 score, is a harmonic mean of precision and recall, providing a comprehensive evaluation of a model's performance, particularly in imbalanced datasets.

$$\text{F-measures} = \frac{2*\text{Recall}*\text{Precision}}{\text{Recall} + \text{Precision}}$$

Results

In Table 1 analysis reveals notable sex-specific differences in tumor genomics and patient outcomes. Female patients exhibit a higher fraction of their genome altered (0.207488) compared to male patients (0.181764), indicating more extensive genomic changes in female tumors. Conversely, male patients present with a higher mutation count (7.435573) and a higher nonsynonymous tumor mutation burden (TMB) (7.305357) than female patients (6.846098 and 6.790197, respectively). Despite these differences, overall survival is similar between the sexes, with female patients showing a marginally longer survival (12.505943 months) compared to male patients (12.406399 months). Tumor purity is nearly identical, with female and male patients having values of 45.528834 and 45.726689, respectively.

Table 2 examines tumor genomic characteristics and survival based on smoking history. Never-smokers have a fraction of genome altered of 0.197811, a mutation count of 6.076468, and a nonsynonymous tumor mutation burden (TMB) of 5.981001. In contrast, previous or current smokers have a slightly lower fraction of genome altered (0.191291) but a significantly higher mutation count (8.144559) and TMB (8.067902). Those with unknown smoking history have intermediate values, with a fraction of genome altered at 0.194683, mutation count of 7.489166, and TMB of 7.364490. Overall survival is highest among never-smokers (13.418332 months), followed by previous or current smokers (12.821535 months), and lowest for patients with unknown.

Table 3 results provide insights into the significance and impact of various predictors on the outcome variable (metastasis). The coefficients (coef) and their exponential values (exp(coef)) indicate the direction and magnitude of the relationship between each predictor and the outcome. Among the predictors, Smoking History and Fraction Genome Altered emerged as significant factors. Smoking History has a positive coefficient of 0.13 with an associated z-value of 5.18 and a p-value less than 0.005, indicating a strong positive association with the outcome, as those with a history of smoking have an estimated 14% increase in the odds of the outcome occurring. Fraction Genome Altered shows a substantial positive coefficient of 0.85, with an exp(coef) of 2.34, suggesting that higher genomic alterations are associated

Table 1 Sex-Specific Differences in Tumor Genomics and Clinical Outcomes

Sex	Fraction Genome Altered	Mutation Count	Overall Survival (Months)	TMB (nonsynonymous)	Tumor Purity
Female	0.207488	6.846098	12.505943	6.790197	45.528834
Male	0.181764	7.435573	12.406399	7.305357	45.726689

TMB=tumor mutation burden

Table 2 Differences in Tumor Genomics and Clinical Outcomes by Smoking History

Smoking History	Fraction Genome Altered	Mutation Count	Overall Survival (Months)	TMB (nonsynonymous)	Tumor Purity
Never	0.197811	6.076468	13.418332	5.981001	46.680397
Prev/Curr Smoker	0.191291	8.144559	12.821535	8.067902	43.823368
Unknown	0.194683	7.489166	9.126138	7.364490	47.192308

TMB=tumor mutation burden

with significantly increased odds (2.34 times) of the outcome, supported by a high z-value of 9.66 and a $p < 0.005$.

In contrast, cancer type, cancer type detailed, and primary tumor site had minimal impact, with coefficients near zero. While the primary tumor site had a statistically significant z-value (−2.56), its effect size was negligible (coef ≈ −0.00). Sex and mutation count were not significant predictors (p=0.45 and 0.74, respectively), indicating limited influence on metastatic outcomes.

Table 3 Significance and Impact of Predictors on Cancer Metastasis (Cox Proportional Hazards Model)

	coef	Exp (coef)	SE (coef)	Coef 95% CI Lower	Coef 95% CI Upper	Exp (Coef) 95% CI Lower	Exp (Coef) 95% CI Upper	cmp to	z	p	$-\log_2$ (p)
Cancer Type	0.00	1.00	0.00	-0.00	0.00	1.00	1.00	0.00	1.57	0.12	3.10
Cancer Type Detailed	0.00	1.00	0.00	0.00	0.00	1.00	1.00	0.00	2.47	0.01	6.22
Primary Tumor Site	-0.00	1.00	0.00	-0.00	-0.00	1.00	1.00	0.00	-2.56	0.01	6.57
Sex	-0.03	0.97	0.04	-0.10	0.04	0.90	1.05	0.00	-0.76	0.45	1.16
Smoking History	0.13	1.14	0.03	0.08	0.19	1.09	1.20	0.00	5.18	<0.005	22.14
Fraction Genome Altered	0.85	2.34	0.09	0.68	1.02	1.97	2.77	0.00	9.66	<0.005	70.95
Mutation Count	0.00	1.00	0.00	-0.00	0.00	1.00	1.00	0.00	0.34	0.74	0.44

Table 4 Feature Importance in Predicting Cancer Metastasis (Feature Selection Analysis)

Features	Score	Percentage (%)
Cancer Type Detailed	0.135	13.5
Fraction Genome Altered	0.447	44.7
Mutation Count	0.183	18.3
Sex	0.030	3.0
Smoking History	0.045	4.5
Cancer Type	0.068	6.8

These findings underscore the importance of smoking history and the extent of genomic alterations as key factors in the model, while other variables appear to have limited or no significant impact.

Table 4 analysis revealed pivotal insights into the predictors influencing cancer metastasis. Among the examined factors, genomic alterations emerged as the most influential predictor, exhibiting a substantial importance score of 0.447. This underscores the significance of genomic instability in driving cancer progression and metastasis, aligning with existing literature emphasizing the role of genomic alterations in tumor development and spread. Following closely behind, mutation count also demonstrated notable importance (0.183), emphasizing its relevance in predicting metastatic potential. These findings highlight the centrality of genomic instability and mutational burden in dictating cancer progression, offering potential avenues for targeted therapeutic interventions aimed at mitigating metastatic risks.

Additionally, the analysis shed light on the predictive value of specific tumor characteristics. Detailed cancer type information and primary tumor site displayed moderate importance scores (0.135 and 0.089, respectively), suggesting their contributions to metastasis prediction. Broad cancer-type categorizations and smoking history exhibited lower importance scores and still contributed significantly to the predictive model. Conversely, sex emerged as the least influential factor, with minimal impact on metastasis prediction. These results emphasize the multifactorial nature of cancer metastasis, implicating a combination of genetic, environmental, and clinical factors in driving disease progression. Overall, the findings underscore the complexity of metastatic processes and offer valuable insights into potential targets for therapeutic intervention and personalized treatment strategies aimed at mitigating cancer metastasis.

Table 5 Performance of AI-ML Algorithms in Predicting Cancer Metastasis

AI-ML Algorithm	Accuracy Score	Train-Test Split
XGBoost	0.95	70:30
SVM	0.95	70:30
Random Forest (RF)	0.90	70:30
Neural Network (NN)	0.80	70:30

Table 5 compares the performance of four AI/ML models in predicting metastasis. XGBoost and Support Vector Machine (SVM) achieved the highest classification accuracy (0.95) using a 70:30 train-test split, indicating excellent predictive ability. These results indicate robust predictive capabilities for both algorithms, suggesting their effectiveness in accurately classifying metastatic outcomes based on the provided features. Following closely behind, the Random Forest algorithm achieved an accuracy score of 0.90 with the same train-test split ratio. While slightly lower

Table 6 Predictive Accuracy of AI-ML Models Based on Confusion Matrix Analysis

Metastasis Type	Precision	Recall	F1-score	Support
Primary	1.00	0.92	0.96	13
Secondary	8.00	1.00	0.93	7
Accuracy			0.95	20
Macro Avg	0.94	0.96	0.95	20
Weighted Avg	0.96	9.95	0.95	20

than XGBoost and SVM, this score still signifies strong predictive performance, highlighting the efficacy of ensemble learning methods in cancer metastasis prediction tasks. In contrast, Neural Networks demonstrated comparatively lower performance, achieving an accuracy score of 0.80. Despite its neural architecture's complexity and potential for learning intricate patterns, Neural Networks may require further optimization or feature engineering to enhance its predictive capabilities for this specific task.

Overall, the results underscore the effectiveness of XGBoost, SVM, and Random Forest algorithms in accurately predicting cancer metastasis based on the provided features. These findings offer valuable insights into selecting appropriate machine learning models for cancer prognosis and personalized treatment planning, ultimately contributing to improved patient outcomes in clinical settings.

Table 6 provides a classification report, detailing precision, recall, and F1-score for predicting primary and secondary metastases. For the primary metastasis class, the model achieved a precision of 1.00, indicating that when it predicts an instance as primary metastasis, it is almost always correct. The recall, which measures the ability to correctly identify all instances of primary metastasis, is 0.92, suggesting that the model successfully captures a high proportion of actual primary metastasis cases. The F1-score, which balances precision and recall, is 0.96, reflecting overall good performance in predicting primary metastasis. The support value indicates that there are 13 instances of primary metastasis in the dataset.

For the secondary metastasis class, precision was 0.88, implying that there may be some false positive predictions. However, the recall is 1.00, indicating that the model correctly identifies all instances of secondary metastasis. The F1-score for secondary metastasis is 0.93, suggesting a reasonably balanced performance between precision and recall. The support value indicates that there are 7 instances of secondary metastasis. Overall, the model achieves an accuracy of 0.95, meaning that it correctly predicts the metastasis type for 95% of the instances in the dataset.

Discussion

This study provides a comprehensive evaluation of the performance of various AI-ML algorithms in predicting cancer metastasis. XGBoost and Support Vector Machine (SVM) demonstrated the highest accuracy scores of 0.95 in a 70:30 train-test split, indicating their effectiveness in distinguishing between primary and secondary metastasis. These algorithms outperformed RF and NN, which achieved accuracy scores of 0.90 and 0.80, respectively. While XGBoost and SVM showcased superior predictive capabilities, RF, despite slightly lower accuracy, still performed reasonably well. NN, although the least accurate among the models evaluated, might have potential for improvement through hyperparameter tuning or additional data preprocessing. These results are consistent with Tapak at el.[16] who found SVM to outperform other machine learning models—including Naïve Bayes, RF, AdaBoost, Logistic Regression, and Linear Discriminant Analysis—in predicting breast cancer outcomes.

The study also reveals that smoking history significantly impacts tumor genomics and patient outcomes. Patients without smoking have a higher mutation count and TMB, while those with a history of smoking have a lower fraction of genome alteration. Overall survival is highest among people who have never smoked. Sex-specific differences in tumor genomics and patient outcomes shows female patients more extensive genomic changes and higher mutation counts. Despite these differences, overall survival is similar, with female patients having a marginally longer survival time. Tumor purity is nearly identical. Furthermore, the study reveals that smoking history and genomic alterations are significant predictors of metastasis. Smoking history increases the odds of the outcome by 14%, while genomic alterations increase the odds by 2.34 times. This analysis was further done on AI-powered algorithms and results revealed that genomic alterations and mutation count are the most influential predictors of cancer metastasis, with genomic instability driving progression and metastasis. Specific tumor characteristics, such as cancer type and primary tumor site, also contribute to metastasis prediction. However, sex is the least influential factor, compared to smoking history, primary tumor site and fraction genome alteration. These findings highlight the multifactorial nature of cancer metastasis, highlighting the need for targeted therapeutic interventions and personalized treatment strategies to mitigate metastatic risks.

These findings underscore the multifactorial nature of metastasis. Chakraborty et al.[17]

previously demonstrated that prostate cancer patients with a high FGA (>6%) but low mutation count (<30 mutations/case) had shorter disease-free survival. In conclusion, cases with a high fraction of genome altered are related to aggressive illness, and those with lower mutational counts might be related to diminished immune responsiveness. Genomic Alterations add to bigger tumor size, higher tumor evaluation, and receptor pessimism. Unmistakable gatherings of genomic changes were seen as related to various evaluations of Invasive Ductal Carcinomas (IDCs). TP53 change was found to assume a significant job in characterizing high tumor grade and is related to a mutation phenotype.[18]

The clinical implications of mutation count have also been explored in prior studies. Bettegowda et al.[19] identified a circulating tumor DNA (ctDNA)-based model showing that lower mutation counts were associated with better prognosis in late-stage lung adenocarcinoma patients treated with chemoradiation. Furthermore, smoking has been implicated in increased metastatic potential in lung and colorectal cancers. Tseng et al.[20] further alluded that there appears to be a link between cigarette smoking and the development of lung metastatic illness in breast cancer patients. This is also highlighted by Makino et al.[21] that cigarette smoking may contribute to the pathophysiology and development of lung metastasis in CRC by increasing adhesion and inflammation.

In summary, the findings emphasize the critical role of genomic alterations (Fraction Genome Altered) and mutation count, smoking history in predicting cancer metastasis. Detailed cancer type information and the primary tumor site also contribute to the prediction. Sex appears to have the least impact on metastasis, but further studies should be done to further assist in identifying key insights into these variables and interventions in managing cancer metastasis.

The study underscores the pivotal role of AI-powered analysis in enhancing early detection of metastasis and personalizing treatment strategies in cancer care. Utilizing a comprehensive dataset of genomic and clinical factors, the AI algorithms identified key predictors of cancer prognosis, particularly focusing on early metastatic detection. The survival analyses pinpointed fraction genome altered (coef= 0.77, $p<0.005$), primary tumor site (coef=0.00, $p<0.005$), and smoking history (coef=0.14, $p<0.005$) as significant factors influencing metastasis. Among these, fraction genome altered exhibited the highest impact on outcomes (Exp(coef)=2.17), followed by smoking history (Exp(coef) = 1.15), highlighting their critical roles in disease progression. The AI-powered models demonstrated high predictive accuracy, with XGBoost and SVM achieving a predictive power accuracy score of 0.95. These models effectively utilized the detailed cancer type, primary tumor site, fraction genome alteration, and mutation count to predict metastasis rates. RF and NN also showed notable performances, with accuracy scores of 0.90 and 0.80, respectively. These findings reinforce the potential of integrating AI in clinical workflows to improve the early detection and treatment of metastatic cancer.

To enhance early metastatic detection and personalize treatment, integration of high-impact predictors such as FGA, primary tumor site, and smoking history into routine clinical workflows is recommended. Given their high accuracy, AI models like XGBoost and SVM should be prioritized in predictive oncology. Genomic testing to assess FGA and mutation burden can help stratify patients by risk. In parallel, implementing effective smoking cessation programs may reduce metastasis risk and improve outcomes.

In conclusiom, this study highlights the significant role of genomic alterations and mutation counts, along with smoking history, in predicting cancer metastasis. AI-powered algorithms, especially XGBoost and SVM, exhibited high accuracy and offer promise in clinical application. Tumor type and primary site further enhance predictive capability, while sex showed minimal influence. These findings emphasize the multifactorial nature of metastasis and suggest that personalized treatment strategies, based on genomic testing and lifestyle factors such as smoking history, are crucial for improving patient outcomes. The integration of AI-powered models in clinical workflows could greatly enhance early metastatic detection and guide targeted therapeutic interventions, reinforcing the importance of continuous research and collaboration to refine predictive models and enhance cancer care.

References

1. Fares J, Fares MY, Khachfe HH, Salhab HA, Fares Y. Molecular principles of metastasis: a hallmark of cancer revisited. Signal Transduct Target Ther. 2020;5(1):28. doi:10.1038/s41392-020-0134-x
2. Crotti M, Bosio A, Invernizzi PL. Validity and reliability of submaximal fitness tests based on perceptual variables. J Sports Med Phys Fitness. 2018;58(5):555–62. doi:10.23736/S0022-4707.17.07199-7
3. Grisham J., When Cancer Spreads: Research Focuses on Better Ways to Treat Metastasis [Internet]. Memorial Sloan Kettering Cancer Center. 2017 [cited 2024 Aug 7]. Available from: https://www.mskcc.org/news/when-cancer-spreads-research-focuses-better-ways-treat-metastasis.
4. Martin TA, Ye L, Sanders AJ, Lane J, Jiang WG. Cancer Invasion and Metastasis: Molecular and Cellular Perspective [Internet]. Nih.gov. Landes Bioscience; 2013. Available from: https://www.ncbi.nlm.nih.gov/books/NBK164700/.
5. Yang Y, Zheng H, Zhan Y, Fan S. An emerging tumor invasion mechanism about the collective cell migration. Am J Transl Res. 2019;11(9):5301–12.
6. Zhang J, Goliwas KF, Wang W, Taufalele PV, Bordeleau F, Reinhart-King CA, *et al*. Energetic regulation of coordinated leader-follower dynamics during collective invasion of breast cancer cells. Proc Natl Acad Sci U S A. 2019;116(16):7867–72. doi:10.1073/pnas.1809964116
7. Huang D, Sun W, Zhou Y, Li P, Chen F, Chen H, *et al*. Mutations of key driver genes in colorectal cancer progression and metastasis. Cancer Metastasis Rev. 2018;37(3):173–7. doi:10.1007/s10555-017-9726-5
8. Yu L, Huang Z, Xiao Z, Tang X, Zeng Z, Tang X, *et al*. Unveiling the best predictive models for early-onset metastatic cancer: insights and innovations (review). Oncol Rep. 2024;51(4):60. doi:10.3892/or.2024.8719
9. Zehir A, Benayed R, Shah RH, Syed A, Middha S, Kim HR,, *et al*. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat Med. 2017;23(6):703–13. doi:10.1038/nm.4333
10. Cockeran M, Meintanis SG, Allison JS. Goodness-of-fit tests in the Cox proportional hazards model. Commun Stat Simul Comput. 2021;50(12):4132–43. doi:10.1080/03610918.2019.1639738
11. Zebari R, Abdulazeez A, Zeebaree D, Zebari D, Saeed J. A comprehensive review of dimensionality reduction techniques for feature selection and feature extraction. J Appl Sci Technol Trends. 2020;1(1):56–70. doi:10.38094/jastt1224
12. Parmar A, Katariya R, Patel V. A review on random forest: an ensemble classifier. In: Hemanth J, Fernando X, Lafata P, Baig Z, eds. International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018. Lecture Notes on Data Engineering and Communications Technologies. 25 ed. Springer; 2019:758-69. doi:10.1007/978-3-030-03146-6_86
13. Chen T, Guestrin C. XGBoost: a scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery; 2016:785–94. doi:10.1145/2939672.2939785
14. Xiao M, Li Y, Yan X, Gao M, Wang W. Convolutional neural network classification of cancer cytopathology images: taking breast cancer as an example. In: Proceedings of the 2024 7th International Conference on Machine Vision and Applications (ICMVA '24). Association for Computing Machinery; 2024:145–9. doi:10.1145/3653946.3653968
15. Naseer I, Masood T, Akram S, Jaffar A, Rashid M, Iqbal MA. Lung cancer detection using modified AlexNet architecture and support vector machine. Comput Mater Contin. 2023;74(1):2039–54. doi:10.32604/cmc.2023.032927
16. Tapak L, Shirmohammadi-Khorram N, Amini P, Alafachi B, Hamidi O, Poorolajal J. Prediction of survival and metastasis in breast cancer patients using machine learning classifiers. Clin Epidemiol Glob Health. 2019;7(3):293–9. doi:10.1016/j.cegh.2018.10.003
17. Chakraborty G, Ghosh A, Nandakumar S, Armenia J, Mazzu YZ, Atiq MO *et al*. Fraction genome altered (FGA) to regulate both cell autonomous and non-cell autonomous functions in prostate cancer and its effect on prostate cancer aggressiveness. J Clin Oncol. 2020;38(6_suppl):347–7. doi:10.1200/JCO.2020.38.6_suppl.347

18. Stephenson Clarke JR, Douglas LR, Duriez PJ, Balourdas DI, Joerger AC, Khadiullina RJ, *et al*. Discovery of nanomolar-affinity pharmacological chaperones stabilizing the oncogenic p53 mutant Y220C. ACS Pharmacol Transl Sci. 2022;5(11):1169–80. doi:10.1021/acsptsci.2c00164
19. Bettegowda C, Sausen M, Leary RJ, Kinde I, Wang Y, Agrawal N, *et al*. Detection of circulating tumor DNA in early- and late-stage human malignancies. Sci Transl Med. 2014;6(224):224ra24. doi:10.1126/scitranslmed.3007094
20. Tseng YJ, Huang CE, Wen CN, Lai PY, Wu MH, Sun YC, *et al*. Predicting breast cancer metastasis by using serum biomarkers and clinicopathological data with machine learning technologies. Int J Med Inform. 2019;128:79-86. doi:10.1016/j.ijmedinf.2019.05.003
21. Makino A, Tsuruta M, Okabayashi K, Ishida T, Shigeta K, Seishima R, *et al*. The impact of smoking on pulmonary metastasis in colorectal cancer. Onco Targets Ther. 2020;13:9623–9. doi:10.2147/OTT.S263250