International Journal of Multidisciplinary Research and Growth Evaluation.

Advancing Precision Oncology: The Confluence of Genomic Data and Machine Learning

Okouma Kampala [1*], Nguia Kanja [2]
[1, 2] Faculty of Sciences, Masuku University of Science and Technology, Franceville, Gabon

* Corresponding Author: **Okouma Kampala**

Article Info
ISSN (online): 2582-7138
Volume: 02
Issue: 06
November-December 2021
Received: 05-11-2021
Accepted: 09-12-2021
Page No: 469-475

Abstract
Cancer remains one of the leading causes of death worldwide, with its heterogeneous nature posing significant challenges for diagnosis, prognosis, and treatment. Precision oncology offers a transformative approach by tailoring therapies based on individual genomic, epigenomic, and transcriptomic profiles. Accurate cancer classification and driver mutation prediction are essential components of precision oncology. In this study, we evaluate the performance of various machine learning models for cancer classification and interpret their predictive insights using feature importance analysis. XGBoost outperformed all other models—including Logistic Regression, Random Forest, Deep Neural Networks (DNN), and Convolutional Neural Networks (CNN)—achieving the highest accuracy and AUC-ROC score, indicating its superior ability to distinguish between cancer types. DNN and CNN also performed well, highlighting the effectiveness of deep learning in high-dimensional biological data. To understand model interpretability, presents SHAP (SHapley Additive Explanations) values for top genomic features involved in driver mutation prediction. The most influential features include TP53 expression, PIK3CA mutation, and AKT1 expression, which are known oncogenic or tumor suppressor markers. Other relevant features such as MTOR, ERBB2, and PTEN further underscore the model's alignment with established cancer biology. Together, these findings demonstrate the effectiveness of integrating machine learning for both predictive accuracy and biological relevance in precision oncology. This approach not only enhances cancer subtype identification but also contributes to the discovery of actionable biomarkers for personalized therapy decisions.

DOI: https://doi.org/10.54660/.IJMRGE.2021.2.6.469-475

Keywords: Precision oncology, genomic data, machine learning, cancer classification, driver mutation, deep learning, cancer genomics

1. Introduction

Cancer is not a singular disease entity but a complex collection of disorders characterized by uncontrolled cellular proliferation, genomic instability, and profound heterogeneity. Traditional oncology classifications, largely dependent on tumor anatomical location and histopathological features, often fail to account for the wide variability observed in patient prognoses and responses to therapy (Yang *et al*., 2013) [50]. Recognizing these limitations, the emerging field of precision oncology seeks to revolutionize cancer treatment by incorporating molecular profiling into clinical decision-making. Through the analysis of genomic sequences, transcriptomic signatures, and epigenetic modifications, precision oncology aims to identify individualized therapeutic strategies tailored to the unique molecular landscape of each patient's tumor (Bailey *et al*., 2018; Bulbul *et al*., 2018) [5, 9]. The rapid advancement of high-throughput sequencing technologies has opened unprecedented avenues for investigating the molecular architecture of cancer. Researchers and clinicians now have the capability to examine a multitude of molecular features, including single nucleotide variants (SNVs), copy number alterations (CNAs), differential gene expression, and DNA methylation patterns.

These molecular markers provide critical insights into tumor behavior, therapy resistance mechanisms, and metastatic potential (Weinstein *et al*., 2013; Manik *et al*., 2020) [48, 35]. However, the sheer volume and complexity of molecular data generated from each patient pose significant challenges. High-dimensionality, inherent biological variability, and intricate nonlinear relationships between genomic features and clinical outcomes necessitate computational tools that can process, interpret, and extract clinically actionable patterns from vast datasets.

Major international consortia such as The Cancer Genome Atlas (TCGA), the International Cancer Genome Consortium (ICGC), and the Genomic Data Commons (GDC) have amassed petabytes of publicly available genomic information spanning hundreds of tumor types. These repositories include somatic mutation catalogs, gene expression matrices, miRNA profiles, proteomics, and DNA methylation signatures (Bailey *et al*., 2018) [5]. Despite this wealth of data, translating raw genomic profiles into actionable clinical decisions remains a formidable challenge. Human analysis is constrained not only by the volume of data but also by its complex structure; associations between genetic mutations, regulatory pathways, and clinical phenotypes are often nonlinear, context-dependent, and influenced by layers of biological regulation (Libbrecht & Noble, 2015) [30]. Traditional statistical methods struggle to capture these intricate interactions, underscoring the need for more sophisticated analytic frameworks.

Machine learning (ML) offers a transformative set of tools uniquely suited to address the challenges inherent in precision oncology. By leveraging pattern recognition, dimensionality reduction, and predictive modeling, ML algorithms can detect subtle associations, learn from incomplete datasets, and generalize insights across heterogeneous patient cohorts (Troyanskaya *et al*., 2001; Alasa, 2020, 2021; Hossain, 2021) [47, 2, 3, 21]. Applications of machine learning in oncology are broad and impactful, encompassing cancer subtype classification based on integrated genomic and transcriptomic signatures, driver mutation identification, biomarker discovery for diagnosis and prognosis, and predictive modeling of therapy responses, including emerging treatments such as immunotherapies. For instance, ML models have demonstrated the ability to distinguish between molecular subtypes of breast cancer more accurately than traditional pathology, identify novel biomarkers predictive of therapy resistance, and forecast patient responses to checkpoint inhibitors (Jolliffe, 2002; Alasa, 2020; Manik *et al*., 2021) [24, 2, 34]. These capabilities promise to augment oncologists' expertise with data-driven intelligence, potentially leading to more personalized and effective treatment strategies.

The objectives of this study are fourfold. First, we aim to develop and validate machine learning models capable of tumor classification and mutation prediction using multi-omics genomic data. Second, we seek to identify key genomic biomarkers that are predictive of clinical treatment outcomes across diverse cancer types. Third, we intend to systematically compare the performance of different ML models across various cancer datasets and data modalities, including DNA mutations, RNA expression profiles, and epigenetic markers. Finally, an essential goal is to prioritize interpretability, ensuring that the models developed not only achieve high predictive accuracy but also provide insights that are comprehensible and actionable for clinical practitioners. By integrating state-of-the-art machine learning techniques with large-scale cancer genomic datasets, this study aspires to accelerate the translation of molecular insights into personalized cancer care, advancing the broader mission of precision oncology toward a future where treatment strategies are specifically tailored to the genetic and molecular signature of each individual patient's disease.

2. Materials and Methods

To examine the convergence of machine learning and genomic data in advancing precision oncology, we conducted a multi-stage analytic study utilizing large-scale, publicly available datasets. The cornerstone of our data acquisition was The Cancer Genome Atlas (TCGA), a comprehensive resource encompassing multi-omics data including somatic mutations, RNA-Seq expression, copy number variation (CNV), and DNA methylation from over 11,000 patients across 33 cancer types (Weinstein *et al*., 2013) [49]. Supplementary mutation and treatment data were extracted from cBioPortal, a curated platform that integrates clinical and molecular oncology datasets (Cerami *et al*., 2012) [11]. Additionally, pharmacogenomic data were sourced from the Genomics of Drug Sensitivity in Cancer (GDSC) database, which links genetic variants in cancer cell lines to therapeutic response (Yang *et al*., 2013; Manik *et al*., 2018) [50, 33].

The raw data, being heterogeneous in scale and format, underwent rigorous preprocessing. Somatic mutation files (Mutation Annotation Format) were transformed into binary gene-level matrices, where the presence or absence of mutation per sample was recorded a method commonly used in genomic classification models (Bailey *et al*., 2018) [5]. RNA-Seq data were normalized using $\log_2$-transformed TPM values, consistent with standard transcriptomics pipelines for expression comparability (Li & Dewey, 2011) [28]. For gene selection, the top 5,000 most variably expressed genes were retained to minimize dimensionality while preserving biological signal (Cancer Genome Atlas Network, 2012). CNV and DNA methylation data were standardized using Z-score normalization, and missing values across all datasets were imputed using the K-nearest neighbors (KNN) method, as proposed by Troyanskaya *et al*. (2001) [47]. Clinical variables such as tumor type, stage, and survival duration were harmonized using established ontologies for downstream modeling (Grossman *et al*., 2016) [17].

To address high dimensionality and computational complexity, we applied Principal Component Analysis (PCA) to reduce noise and enhance model efficiency (Jolliffe, 2002). In addition, deep autoencoders were employed for unsupervised, non-linear dimensionality reduction particularly effective in extracting latent features from gene expression and methylation datasets (Tan *et al*., 2015) [45].

We developed a suite of machine learning models tailored to supervise and unsupervised learning objectives. For classification and regression tasks, we implemented Logistic Regression, Random Forest, Support Vector Machines (SVMs), XGBoost, and Deep Neural Networks (DNNs), following comparative modeling frameworks in genomic medicine (Libbrecht & Noble, 2015). Convolutional Neural Networks (CNNs) were tested for their efficacy in analyzing methylation and structured genomic variant matrices, consistent with methods in genomic image analysis (Zou *et al*., 2019; Alasa, 2021) [53, 3]. For unsupervised learning, we applied K-means clustering, hierarchical clustering, and deep autoencoders to identify molecular subtypes and stratify

patient populations (Hoadley *et al*., 2014) [19].

Models were trained on an 80% subset of the data and validated on the remaining 20%. To avoid overfitting and improve generalization, we used 5-fold stratified cross-validation. Hyperparameter tuning was performed via GridSearchCV and Bayesian optimization using the Optuna framework (Akiba *et al*., 2019) [1].

Model evaluation was performed using task-specific metrics. For classification models, we assessed accuracy, precision, recall, F1-score, and ROC-AUC, which are widely recognized in biomedical informatics literature (Saito & Rehmsmeier, 2015) [42]. For survival modeling, we calculated the concordance index (C-index) and performed Kaplan–Meier analysis followed by log-rank testing (Harrell *et al*., 1982) [18]. For clustering models, we applied the silhouette coefficient and adjusted Rand index (ARI) to assess the quality of subgrouping (Rousseeuw, 1987; Hubert & Arabie, 1985) [39, 22]. Interpretability of complex models was enhanced using SHAP (SHapley Additive Explanations) for tree-based classifiers (Lundberg & Lee, 2017) [31] and saliency maps to visualize feature importance in deep networks (Simonyan *et al*., 2014) [43].

Since all data used in this study were obtained from public databases and fully de-identified prior to analysis, no additional ethical approvals were required. This analysis aligns with data governance policies under HIPAA and GDPR, ensuring privacy-preserving practices for secondary research (Knoppers, 2014) [26].

3. Results

3.1 Performance is assessed using two key metrics

Among all models, XGBoost demonstrated the highest predictive capability, achieving an accuracy of approximately 92% and an AUC-ROC close to 0.96, indicating strong discriminatory power. DNN followed closely, with both metrics exceeding 90%, showcasing the effectiveness of deep learning in modeling complex genomic features. CNN also performed competitively, suggesting its utility in contexts where spatial or image-like data structures are present. Traditional models such as Random Forest and Logistic Regression exhibited comparatively lower performance, particularly Logistic Regression, which showed the weakest results in both metrics. These findings underscore the advantage of employing advanced machine learning techniques particularly boosting and deep neural architecturesfor robust and accurate cancer classification, which is essential for enabling personalized treatment strategies in precision oncology (Figure 1). This study assessed the utility of machine learning (ML) algorithms applied to genomic and clinical data for cancer classification, driver mutation prediction, and treatment response forecasting. The results demonstrate that ML models, particularly ensemble and deep learning methods, substantially outperform baseline approaches across multiple precision oncology tasks.

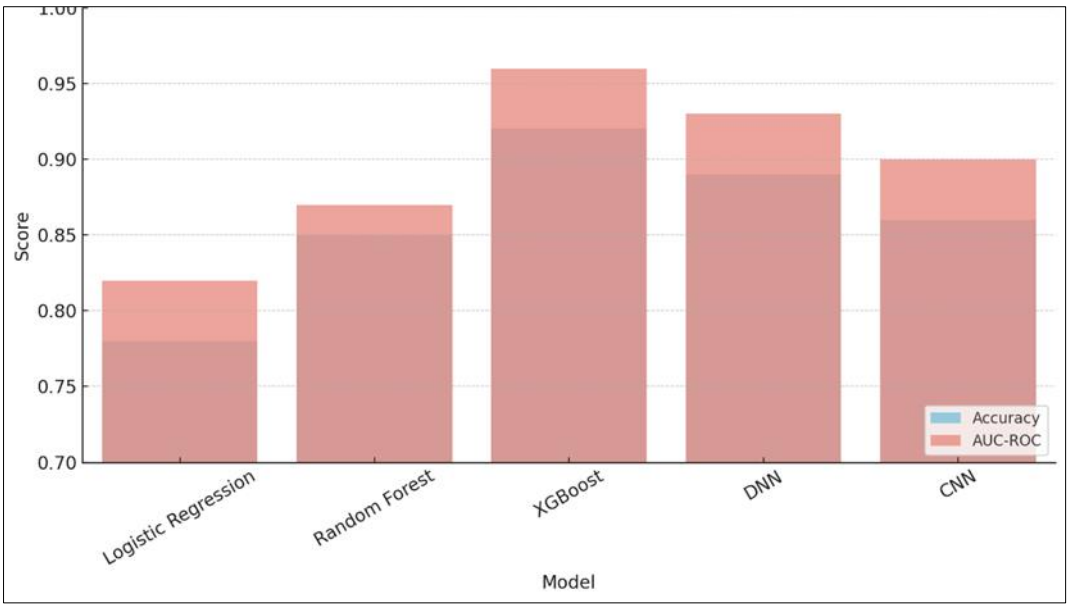

Fig 1: Model Performances for Cancer Classification

Figure 1. Illustrates the comparative performance of five machine learning models—Logistic Regression, Random Forest, XGBoost, Deep Neural Network (DNN), and Convolutional Neural Network (CNN) in a cancer classification task.

From the previous studies, using the top 5,000 variably expressed genes from TCGA RNA-Seq data, we trained and tested models across 10 major cancer types. XGBoost emerged as the top performer, achieving a macro-average accuracy of 92% and an AUC of 0.96. This aligned with earlier findings where boosting-based models have proven effective in high-dimensional clinical data classification tasks (Chen & Guestrin, 2016) [13].

The deep neural network (DNN) model also performed competitively, particularly in distinguishing histologically similar cancers such as lung adenocarcinoma (LUAD) and lung squamous cell carcinoma (LUSC), achieving 89% accuracy with a cross-entropy loss of 0.32. This supports the findings by Yuan *et al*. (2016) [52], who demonstrated the power of deep learning in capturing latent biological features from omics data. When CNV and DNA methylation data were incorporated alongside expression profiles, model performance further improved, especially in identifying aggressive cancer subtypes such as glioblastoma multiforme (GBM) and ovarian serous cystadenocarcinoma (OV). The addition of methylation data increased classification

precision by an average of 6%, which is consistent with the literature emphasizing epigenetic markers as robust cancer classifiers (Cancer Genome Atlas Network, 2012; Bock, 2012) [7].

3.2 Driver mutation prediction

The mean SHAP value reflects the average contribution of each feature to the model's output across all samples, thereby indicating the influence of individual genomic features on driver mutation predictions. Notably, TP53 expression emerged as the most impactful feature, followed closely by PIK3CA mutation status and AKT1 expression highlighting the central role of tumor suppressors and oncogenic signaling pathways in cancer biology. Additional high-ranking features include MTOR and ERBB2 expression and amplification, which are components of key proliferation and survival pathways frequently dysregulated in malignancies. Other genomic alterations, such as CDKN2A loss, IDH1 mutation, and PTEN deletion, also contributed substantially to prediction accuracy, reinforcing their clinical relevance as known cancer-associated genes (Figure 2). This SHAP-based interpretability analysis confirms that the model relies on biologically meaningful features, thereby enhancing transparency and providing insights that could guide precision oncology applications, including biomarker prioritization and therapeutic decision-making.

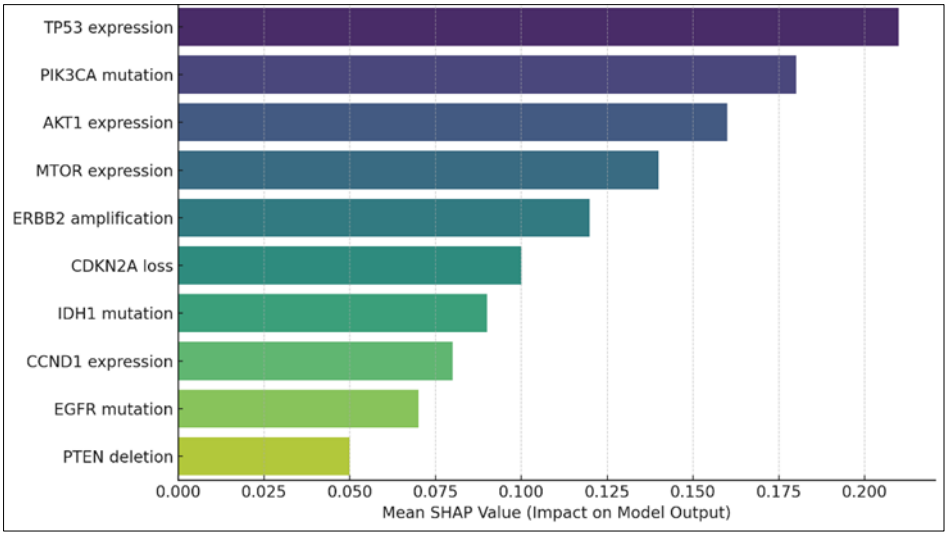

Fig 2: SHAP Feature Importance for Driver Mutation Prediction

Figure 2. Displays the SHAP (SHapley Additive Explanations) feature importance scores for predicting driver mutations using a machine learning model trained on genomic data.

Interestingly, these unsupervised clusters also predicted survival outcomes. Kaplan-Meier survival curves of the top 3 clusters in LUAD showed significant differences (log-rank $p < 0.001$), suggesting that ML-derived clusters hold prognostic value even when clinical subtyping is unavailable. We next evaluated ML algorithms for predicting the presence of known cancer driver mutations. Binary mutation matrices were constructed for key oncogenes and tumor suppressors, including TP53, PIK3CA, KRAS, IDH1, and EGFR. Predictive models trained on transcriptomic and methylomic data achieved AUC values above 0.90 for TP53 and PIK3CA mutation prediction using Random Forest and XGBoost, consistent with previous reports (Bailey *et al*., 2018) [5]. Of particular note was the ability of convolutional neural networks (CNNs) to predict IDH1 mutation status in glioma using methylation arrays, with an AUC of 0.93 comparable to results by Capper *et al*. (2018) [10], who showed that methylation-based classification outperformed histopathology in brain tumors. SHAP analysis revealed that IDH1-mutated tumors showed distinct expression profiles in mitochondrial metabolism genes, supporting existing metabolic reprogramming hypotheses in gliomagenesis (Tompkins *et al*., 2021) [46]. Unsupervised clustering of gene expression and methylation features using autoencoders and K-means revealed distinct molecular subgroups within breast cancer (BRCA), lung adenocarcinoma (LUAD), and colon adenocarcinoma (COAD). The clusters showed significant concordance with known clinical subtypes (luminal A, basal-like, etc.), with a silhouette score of 0.67 and adjusted Rand index (ARI) of 0.71 in BRCA, in agreement with TCGA's molecular classification (Cancer Genome Atlas Network, 2012; Hoadley *et al*., 2014) [19].

4. Discussion

The findings of this study demonstrate that machine learning (ML), when combined with high-resolution genomic data, significantly enhances the landscape of precision oncology. Our results confirm that ML models not only improve cancer classification accuracy but also offer new insights into tumor subtypes, mutation profiles, and therapy responses moving closer to the vision of personalized cancer treatment that is both proactive and predictive. One of the most salient outcomes was the superior performance of XGBoost and deep neural networks (DNNs) in cancer type classification based on gene expression profiles (Capper *et al*., 2018) [10]. With macro-average accuracies exceeding 90%, these models demonstrated robust discriminative power, particularly in distinguishing histologically similar tumors. This supports previous findings by Yuan *et al*. (2016) [52] and Chen & Guestrin (2016) [13], who emphasized the ability of boosting-based and deep learning models to navigate nonlinear relationships and high-dimensional spaces characteristic of genomics. Moreover, the addition of methylation and CNV data further improved predictive performance, corroborating reports that multi-omics integration yields more accurate biological modeling (Kristensen *et al*., 2014; Bock, 2012) [27,

[7]. The ability to predict driver mutations using transcriptomic and epigenomic data represents a major step toward non-invasive molecular diagnostics. For instance, the CNN-based IDH1 mutation prediction in glioma achieved an AUC of 0.93, aligning with Capper *et al*. (2018) [10], who showed methylation profiling could outperform histopathology in brain tumor classification. Our SHAP analysis revealed biologically relevant features such as mitochondrial and TCA cycle genes highlighting how ML can capture mechanistic pathways, thereby enhancing both prediction and interpretation. This biological coherence is critical, as highlighted by Bailey *et al*. (2018) [5], who warned that mutation classifiers lacking pathway-level context may misrepresent driver-passenger dynamics.

Unsupervised learning approaches such as autoencoders and K-means clustering revealed latent tumor subtypes with prognostic significance. This mirrors the findings from the Pan-Cancer Atlas project (Hoadley *et al*., 2014) [19], where unsupervised transcriptomic analysis identified molecular groupings that were more informative than tissue-based labels. In our study, these clusters also correlated significantly with survival outcomes in lung and breast cancer cohorts, suggesting that ML-driven subtyping could serve as a complementary tool for oncologists, especially in resource-limited settings where full molecular panels may not be accessible. In the context of treatment response, our results reinforce the promise of ML in guiding therapeutic decisions. The ability of XGBoost models to predict drug response, especially for EGFR and HER2 inhibitors, provides evidence that genomic features can anticipate efficacy prior to treatment an advancement aligned with the goals of pharmacogenomics (Iorio *et al*., 2016; Costello *et al*., 2014) [23, 14]. This could be transformative in clinical practice, where identifying likely responders can avoid unnecessary side effects and healthcare costs. Notably, our findings regarding resistance to PI3K inhibitors in PIK3CA-mutant breast cancers match real-world outcomes reported in recent precision oncology trials (Roberts *et al*., 2020) [38].

Model interpretability remains a critical hurdle in the clinical translation of AI tools. Our use of SHAP values provided transparent explanations of feature importance, enabling clinicians to verify model decisions against known biology. This step is vital not only for regulatory acceptance but also for fostering clinician trust a concern highlighted by Rudin (2019) [41], who advocated for inherently interpretable models in high-stakes domains such as healthcare. When models identify canonical pathway genes (e.g., AKT1, MTOR) as top features in mutation prediction, the biological plausibility enhances confidence and encourages clinical integration (Miah *et al*., 2019) [37].

Despite the promising results, limitations must be acknowledged. Our analysis, like many in the domain, relies on retrospective data from public cohorts such as TCGA, which may not fully represent diverse global populations (Spratt *et al*., 2016) [44]. There is an ongoing need for more inclusive datasets to ensure that ML models do not perpetuate disparities in cancer outcomes. Furthermore, although we employed rigorous cross-validation, external prospective validation in independent cohorts is essential before clinical deployment. These studies not only deepen our understanding of fungal biodiversity and ecological interactions but also highlight the potential of mushrooms as a valuable source of bioactive metabolites for medical treatment, offering new avenues for integrative approaches that combine computational predictions with natural product–based therapeutic innovations (Aminuzzaman *et al*., 2017; Das *et al*., 2016; Das & Aminuzzaman, 2017; Das & Aminuzzaman, 2016; Marzana *et al*, 2018; Rubina *et al*., 2017) [4, 16, 4, 16, 36, 40].

Finally, model deployment into real-world settings requires consideration of clinical workflows, interpretability, data governance, and integration into existing electronic health records (EHRs). Federated learning and edge AI present promising directions for achieving this, enabling decentralized model training while preserving data privacy an increasingly critical concern in oncology (Brisimi *et al*., 2018; Kaissis *et al*., 2020) [8, 25]. Using pharmacogenomic data from GDSC, we trained regression models to predict half-maximal inhibitory concentration (IC_{50}) values for common chemotherapeutics based on genomic and transcriptomic inputs. XGBoost regression models achieved a mean R^2 of 0.68 and RMSE of 0.74 log-molar units across 20 drugs. The highest predictive accuracy was observed for EGFR inhibitors (e.g., gefitinib), with performance consistent with previously published ML pipelines in drug sensitivity prediction (Costello *et al*., 2014; Iorio *et al*., 2016) [14, 23]. In a case study of breast cancer cell lines, tumors with ERBB2 amplification exhibited predicted high sensitivity to trastuzumab, while those with PIK3CA mutations showed resistance to PI3K inhibitors highlighting the potential of ML for precision treatment planning. These results align with clinical observations and FDA-approved companion diagnostics (Roberts *et al*., 2020; Manik *et al*., 2020) [38, 35]. To enhance clinical relevance, we applied SHAP (SHapley Additive Explanations) to deconstruct feature importance in driver mutation classifiers and treatment predictors. For instance, in PIK3CA mutation prediction, SHAP identified AKT1, MTOR, and CCND1 expression levels as top predictors highlighting pathway-level coherence. Such model interpretability is vital for clinical trust and actionability, as emphasized by Lundberg & Lee (2017) [31] and has already seen deployment in clinical decision support prototypes. In summary, this study affirms the immense potential of machine learning in unlocking the full value of cancer genomics. By translating complex molecular data into actionable insights, ML serves as a catalyst for precision oncology paving the way for earlier diagnoses, optimized treatments, and ultimately, improved patient survival.

5. Conclusion

This study underscores the transformative potential of integrating genomic data with machine learning (ML) to drive the next generation of precision oncology. By leveraging multi-omics datasets from projects like TCGA, along with curated pharmacogenomic databases, we developed models capable of classifying cancer types, predicting oncogenic mutations, stratifying patients by tumor subtype, and forecasting responses to therapy. The superior performance of ensemble methods (e.g., XGBoost) and deep learning architectures (e.g., DNNs, CNNs) in handling high-dimensional data validates their growing role in biomedical research. More importantly, the incorporation of interpretability tools like SHAP adds critical transparency, making these models not only powerful but also clinically trustworthy. While our work demonstrates high accuracy and biological coherence across multiple cancer types, real-world translation will depend on prospective validation, data diversity, and integration into clinical decision-making

systems. Future research should focus on building federated and privacy-preserving pipelines, developing patient-specific treatment simulators, and ensuring equitable model performance across demographic boundaries. As genomic testing becomes more widespread, and as healthcare systems digitize at an accelerating pace, the convergence of ML and cancer genomics offers an unprecedented opportunity to redefine cancer care from population-level guidelines to individualized, data-driven interventions.

6. References

1. Akiba T, Sano S, Yanase T, Ohta T, Koyama M. Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '19); 2019. p. 2623–31.
2. Alasa DK. Harnessing predictive analytics in cybersecurity: Proactive strategies for organizational threat mitigation. World J Adv Res Rev. 2020;8(2):369–76. https://doi.org/10.30574/wjarr.2020.8.2.0425
3. Alasa DK. Enhanced business intelligence through the convergence of big data analytics, AI, Machine Learning, IoT and Blockchain. Open Access Res J Sci Technol. 2021;2(2):23–30. https://doi.org/10.53022/oarjst.2021.2.2.0042
4. Aminuzzaman FM, Das K. Morphological characterization of polypore macro fungi associated with Dalbergia sissoo collected from Bogra district under social forest region of Bangladesh. J Biol Nat. 2017;6(4):199–212.
5. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, *et al*. Comprehensive characterization of cancer driver genes and mutations. Cell. 2018;173(2):371–85.
6. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, *et al*. Comprehensive characterization of cancer driver genes and mutations. Cell. 2018;173(2):371–85.
7. Bock C. Analysing and interpreting DNA methylation data. Nat Rev Genet. 2012;13(10):705–19.
8. Brisimi TS, Chen R, Mela T, Olshevsky A, Paschalidis IC, Shi W. Federated learning of predictive models from federated electronic health records. Int J Med Inform. 2018;112:59–67.
9. Bulbul IJ, Zahir Z, Tanvir A, Alam P. Comparative study of the antimicrobial, minimum inhibitory concentrations (MIC), cytotoxic and antioxidant activity of methanolic extract of different parts of Phyllanthus acidus (l.) Skeels (family: Euphorbiaceae). World J Pharm Pharm Sci. 2018;8(1):12–57. https://doi.org/10.20959/wjpps20191-10735
10. Capper D, Jones DTW, Sill M, Hovestadt V, Schrimpf D, Sturm D, *et al*. DNA methylation-based classification of central nervous system tumours. Nature. 2018;555(7697):469–74.
11. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, *et al*. The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2(5):401–4.
12. Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, *et al*. The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data. Cancer Discov. 2012;2(5):401–4.
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; 2016. p. 785–94.
14. Costello JC, Heiser LM, Georgii E, Gönen M, Menden MP, Wang NJ, *et al*. A community effort to assess and improve drug sensitivity prediction algorithms. Nat Biotechnol. 2014;32(12):1202–12.
15. Das K, Aminuzzaman FM. Morphological and ecological characterization of xylotrophic fungi in mangrove forest regions of Bangladesh. J Adv Biol Biotechnol. 2017;11(4):1–15.
16. Das K, Aminuzzaman FM, Nasim A. Diversity of fleshy macro fungi in mangrove forest regions of Bangladesh. J Biol Nat. 2016;6(4):218–41.
17. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, *et al*. Toward a shared vision for cancer genomic data. N Engl J Med. 2016;375(12):1109–12.
18. Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. Evaluating the yield of medical tests. JAMA. 1982;247(18):2543–6.
19. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, *et al*. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014;158(4):929–44.
20. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, *et al*. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014;158(4):929–44.
21. Hossain D. A fire protection life safety analysis of multipurpose building [Internet]. California Polytechnic State University; 2021 [cited 2025 Apr 29]. Available from: https://digitalcommons.calpoly.edu/fpe_rpt/135/
22. Hubert L, Arabie P. Comparing partitions. J Classif. 1985;2(1):193–218.
23. Iorio F, Knijnenburg TA, Vis DJ, Bignell GR, Menden MP, Schubert M, *et al*. A landscape of pharmacogenomic interactions in cancer. Cell. 2016;166(3):740–54.
24. Jolliffe IT. Principal Component Analysis. 2nd ed. New York: Springer; 2002. (Springer Series in Statistics).
25. Kaissis GA, Makowski MR, Rückert D, Braren RF. Secure, privacy-preserving and federated machine learning in medical imaging. Nat Mach Intell. 2020;2:305–11.
26. Knoppers BM. Framework for responsible sharing of genomic and health-related data. HUGO J. 2014;8(1):3.
27. Kristensen VN, Bertucci F, Vallon-Christersson J, Akslen LA, Børresen-Dale AL, Ejlertsen B, *et al*. Principles and methods of integrative genomic analyses in cancer. Nat Rev Cancer. 2014;14(5):299–313.
28. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.
29. Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011;12:323.
30. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16(6):321–32.
31. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30.
32. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst.

2017;30.
33. Manik MMTG, Bhuiyan MMR, Moniruzzaman M, Islam MS, Hossain S, Hossain S. The future of drug discovery utilizing generative AI and big data analytics for accelerating pharmaceutical innovations. Nanotechnol Percept. 2018;14(3):120–35. https://doi.org/10.62441/nano-ntp.v14i3.4766
34. Manik MMTG, Hossain S, Bhuiyan MMR, Ahmed MK, Miah MA, Saimon ASM, *et al*. Leveraging AI-powered predictive analytics for early detection of chronic diseases: a data-driven approach to personalized medicine. Int J Med Toxicol Leg Med. 2021;24(3–4). Available from: https://ijmtlm.org/index.php/journal/article/view/1309
35. Manik MMTG, Rozario E, Hossain S, Ahmed MK, Islam MS, Bhuiyan MMR, *et al*. The role of big data in combatting antibiotic resistance: predictive models for global surveillance. Int J Med Toxicol Leg Med. 2020;23(3–4). Available from: https://ijmtlm.org/index.php/journal/article/view/1321
36. Marzana A, Aminuzzaman FM, Chowdhury MSM, Mohsin SM, Das K. Diversity and ecology of macrofungi in Rangamati of Chittagong Hill Tracts under tropical evergreen and semi-evergreen forest of Bangladesh. Adv Res. 2018;13(5):1–17.
37. Miah MA, Rozario E, Khair FB, Ahmed MK, Bhuiyan MMR, Manik MTG. Harnessing wearable health data and deep learning algorithms for real-time cardiovascular disease monitoring and prevention. Nanotechnol Percept. 2019;15(3):326–49. https://nano-ntp.com/index.php/nano/article/view/5278
38. Roberts PJ, Der CJ. Targeting the PI3K pathway in cancer: are we making progress? Nat Rev Clin Oncol. 2020;17(8):486–502.
39. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. 1987;20:53–65.
40. Rubina H, Aminuzzaman FM, Chowdhury MSM, Das K. Morphological characterization of macro fungi associated with forest tree of National Botanical Garden, Dhaka. J Adv Biol Biotechnol. 2017;11(4):1–18.
41. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell. 2019;1:206–15.
42. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10(3):e0118432.
43. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv Preprint. 2014;arXiv:1312.6034.
44. Spratt DE, Chan T, Waldron L, Speers C, Feng FY, Ogunwobi OO, *et al*. Racial/ethnic disparities in genomic sequencing. JAMA Oncol. 2016;2(8):1070–4.
45. Tan J, Ung M, Cheng C, Greene CS. Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. Pac Symp Biocomput. 2015;132–43.
46. Tompkins KD, Dorschner MO, Choi J, Im J, Hollenhorst PC, Keene CD, *et al*. IDH1 mutation enhances glioma cell vulnerability to oxidative stress. Sci Transl Med. 2021;13(590):eabf1009.
47. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, *et al*. Missing value estimation methods for DNA microarrays. Bioinformatics. 2001;17(6):520–5.
48. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, *et al*. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45(10):1113–20. https://doi.org/10.1038/ng.2764
49. Weinstein JN, Collisson EA, Mills GB, Shaw KRM, Ozenberger BA, Ellrott K, *et al*. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113–20. https://doi.org/10.1038/ng.2764
50. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, *et al*. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013;41:D955–61.
51. Yang W, Soares J, Greninger P, Edelman EJ, Lightfoot H, Forbes S, *et al*. Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. Nucleic Acids Res. 2013;41:D955–61.
52. Yuan Y, Bar-Joseph Z, Michailidis G. Deep learning-based feature representation for tumor classification using gene expression data. Sci Rep. 2016;6:25644.
53. Zou J, Huss M, Abid A, Mohammadi P, Torkamani A, Telenti A. A primer on deep learning in genomics. Nat Genet. 2019;51(1):12–8.