Journal of Carcinogenesis
Covering Genesis to Therapeutics

Journal of Carcinogenesis, ISSN: 1477-3163
2025; Vol 24: Issue 4s, https://carcinogenesis.com/

Machine Learning Models for Predicting Carcinogenesis Pathways Using Genomic and Environmental Data

Dr. D. Shamia[1], Bharathidhasan. A[2], Dr. Venkat Ghodke[3], Dr. Yamuna N[4], M. Mahabooba[5], S. Ramya[6]

[1] Associate Professor, V.S.B. College of Engineering Technical Campus, Tamil Nadu, India

Email ID: shamiasathish@gmail.com

[2] Assistant Professor /AIDS, VSB Engineering College, Karur, Tamil Nadu, India

Email ID: annabhar87@gmail.com

[3] Assistant Professor, Department: Electronics and Telecommunication Engineering, AISSMS Institute of Information Technology, Pune, Maharashtra, India

Email.ID: venkatghodke@aissmsioit.org / ORC id: https://orcid.org/0000-0002-1249-9692

[4] Assistant Professor, Department of Information Technology, Hindusthan College of Engineering and Technology

Email.ID: yamunagovind@gmail.com

[5]Assistant Professor (SG), Department of Computer Science and Engineering, Nehru Institute of Engineering and Technology, Coimbatore, Tamil Nadu, India

Email ID: nietmahaboobam@nehrucolleges.com

[6]Assistant Professor, Artificial Intelligence and Machine Learning, M. Kumarasamy College of Engineering, Thalavapalayam, Karur, Tamil Nadu, India

Email ID: ramyalaksha@gmail.com

ABSTRACT

Advances in machine learning (ML) provide a powerful framework for understanding the complex interplay between genetic and environmental factors in carcinogenesis. This work develops and evaluates predictive machine learning models to classify the primary oncogenic pathways that lead to the development of cancer in people. By integrating high-dimensional genomic data (such as somatic mutations, copy number variations, and gene expression profiles) with structured environmental exposure data (such as toxins, lifestyle variables, and somatic mutations), the models discover complex, non-linear relationships that traditional statistical methods frequently miss. We compare ensemble methods like Random Forest and Gradient Boosting to more complex deep neural networks (DNNs) for this multi-class classification task. The models are trained and validated using large-scale cohorts such as The Cancer Genome Atlas (TCGA), and key predictive features are identified using feature significance analysis. Because ML models can accurately predict the primary carcinogenesis routes, the results demonstrate that they are a powerful tool for etiological study. This approach enables a more personalised understanding of cancer aetiology, which is critical for developing precision oncology and targeted prevention strategies.

Keywords: *Machine Learning, Carcinogenesis Pathways, Genomic Data, Environmental Exposure, Predictive Modeling, Precision Oncology.*

How to Cite: Dr. D. Shamia, Bharathidhasan. A, Dr. Venkat Ghodke, Dr. Yamuna N, M. Mahabooba, S. Ramya, (2025) Machine Learning Models for Predicting Carcinogenesis Pathways Using Genomic and Environmental Data, *Journal of Carcinogenesis*, *Vol.24, No.4s*, 966-974

1. INTRODUCTION

Cancer is the result of intricate gene-environment interactions. This work combines genetic and environmental data to develop a machine learning model to predict carcinogenesis pathways, improving precision oncology. The objective is to outperform approaches that rely only on one kind of data.

1.1. The Complexity of Carcinogenesis

Cancer's hallmarks are a well-known illustration of the complex, multi-step evolutionary process—rather than a single event—that leads to cancer. This illness, sometimes referred to as carcinogenesis, is driven by the accumulation of genetic and epigenetic alterations that provide cells specific growth advantages. These modifications are acquired by a combination of internal processes, including genetic predispositions and random replication errors, as well as external environmental selection factors. Determining the exact pathways by which this shift occurs for a particular patient remains a major oncology problem, with important implications for prevention, early detection, and treatment.

1.2. The Part of the Genome

The carcinogenesis blueprint is found in the genome. Extensive genome sequencing studies have highlighted the vast spectrum of somatic mutations seen in malignancies. A key distinction is made between driver mutations, which actively contribute to oncogenesis by providing a proliferative advantage, and passenger mutations, which are functionally neutral hitch-hikers. Resources like the COSMIC (Catalogue Of Somatic Mutations In Cancer) database have classified these fingerprints into distinct patterns associated with a number of factors, such as UV radiation exposure, defects in DNA mismatch repair, or APOBEC enzyme activity. Consequently, a tumor's genomic sequence provides a record of the mutational processes that have been active throughout its development.

1.3. The Environmental Component

Environmental influences often offer the directorial cues, whereas genetics supplies the interior screenplay. Numerous epidemiological studies have shown a high correlation between certain cancer types and external exposures. Aflatoxin (liver cancer), Helicobacter pylori (gastric cancer), Human Papillomavirus (HPV; cervical and oropharyngeal cancer), ultraviolet (UV) radiation (melanoma), and tobacco smoke (associated with lung and bladder cancer) are major environmental carcinogens. These substances can cause long-term inflammation or direct damage to DNA, which speeds up mutagenesis and leaves distinctive marks on the genome known as mutational signatures. However, a significant challenge lies in accurately **quantifying lifetime exposures** for individuals, which often relies on self-reported data from questionnaires and is subject to recall bias and imprecision, making it a complex variable to model.

1.4. The Problem of Integration

Despite a clear understanding that gene-environment interactions (GxE) are pivotal in cancer development, genomic and environmental research have largely progressed in parallel silos. Traditional biostatistical methods, such as logistic regression, often struggle to model the high-dimensionality, heterogeneity, and complex, non-linear interactions inherent in these datasets. While a genome-wide association study (GWAS) might identify a genetic risk variant and an epidemiological study might identify an environmental risk factor, the synergy between a genetic predisposition and an environmental trigger is frequently poorly captured. This integrative gap limits our ability to move from population-level risk associations to personalized predictions of carcinogenic pathways, hindering the development of tailored prevention strategies.

1.5. The Promise of Machine Learning

Machine learning (ML) offers a powerful suite of tools to overcome these limitations. ML algorithms are explicitly designed to learn complex patterns from high-dimensional, heterogeneous data without requiring *a priori* assumptions about the relationships between variables. They are exceptionally well-suited for tasks such as identifying subtle interactions between genetic variants and environmental factors, handling mixed data types (e.g., continuous genomic values and categorical exposure data), and performing robust feature selection to identify the most predictive drivers. Techniques from supervised learning, including ensemble methods and DNNs, present a promising framework for building integrative predictive models that can synthesize disparate sources of information.

1.6. Objectives and Hypothesis

The primary **objective** of this paper is to develop and validate a machine learning model that predicts the dominant carcinogenesis pathway of a tumour by integrating genomic features and environmental exposure data. We posit that the most accurate model will be one that synergistically combines both data types.

Our central **hypothesis** is that an integrative ML model, leveraging both genomic and environmental data, will significantly outperform models trained on either data type alone in classifying the etiological pathway of a cancer sample. We expect that the feature importance metrics from the optimal model will yield biologically and clinically interpretable insights into the key drivers of specific carcinogenesis pathways.

1.7. Paper Structure

This paper's remaining sections are organised as follows: Data sources, preprocessing, feature engineering, the chosen machine learning algorithms, and the assessment framework are all covered in depth in Section 2. The outcomes of our comparative performance study, hyperparameter optimisation, and model training are shown in Section 3. The ramifications of our results are discussed in Section 4, along with an interpretation of the key predictive characteristics, an acknowledgement of the study's limitations, and recommendations for future research areas. A final synopsis of the study and its possible contributions to precision oncology is given in Section 5.

2. METHODOLOGY

2.1. Data Acquisition and Curation

- **Genomic Data Source:** Genomic, transcriptomic, and clinical data were downloaded from the Genomic Data Commons (GDC) for cohorts within The Cancer Genome Atlas (TCGA) program. We concentrated on cancer types with known etiological links, such as Lung Adenocarcinoma (LUAD), Lung Squamous Cell Carcinoma (LUSC), Skin Cutaneous Melanoma (SKCM), Liver Hepatocellular Carcinoma (LIHC), and Head and Neck Squamous Cell Carcinoma (HNSC), in order to guarantee strong signal detection for environmental associations.
- **Genomic Features:** We identified four main feature kinds for every tumour sample:
    1. **Mutational Profiles:** The number of minor insertions/deletions (Indels) and non-synonymous Single Nucleotide Variants (SNVs) in each sample.
    2. **Mutational Signatures:** The deconstructSigs R program was used to derive the activity of mutational signatures from each sample's SNV profiles. SBS1 (ageing clock), SBS2/13 (APOBEC activity), SBS4 (tobacco smoking), SBS5 (clock-like), SBS6 (MMR-deficiency), SBS7a/7b/7c (UV light exposure), and SBS22 (aristolochic acid) were the predetermined set of COSMIC v3.2 signatures from which we concentrated.
    3. **Copy Number Variations (CNVs):** Features for amplifications and deletions were included in the focal and arm-level CNV data (GISTIC2.0 calls).
- **Environmental Data Source:** Given the challenge of obtaining direct lifetime exposure metrics, we derived proxy measures from the curated TCGA clinical metadata:
    - **Smoking:** Pack-years and status (Never, Current, Former).
    - **Alcohol Use:** Documented history (Yes/No).
    - **BMI:** Used as a proxy for diet and metabolic factors.
    - **Viral Status:** HPV (p16 status from HNSC) and HBV/HCV (serology from LIHC).
    - **Age at Diagnosis:** A proxy for cumulative endogenous mutational processes.

Label Generation (The Prediction Target): Each sample was assigned a primary "carcinogenesis pathway" label based on a deterministic algorithm using the strongest genomic and clinical evidence:

- **UV-Associated:** SKCM samples with high SBS7 exposure.
- **Tobacco-Associated:** LUAD/LUSC samples with high SBS4 exposure and/or significant smoking history.
- **Viral (HPV):** HNSC samples positive for p16.
- **Viral (HBV/HCV):** LIHC samples with positive serology.
- **MMR-Deficient:** Samples with high SBS6 exposure and/or MSI-H status.
- **APOBEC-Associated:** Samples with dominant SBS2/13 exposure without a stronger primary label.
- **Aging / Endogenous:** Samples with dominant SBS1/SBS5 exposure and no strong environmental driver. Samples that could not be confidently assigned were labeled "Unknown" and excluded from the supervised learning task.

2.2. Data Preprocessing and Feature Engineering

Environmental feature missing data (e.g., pack-years) was filled in using mode imputation for categorical variables and multivariate imputation by chained equations (MICE) for continuous variables. Data on gene expression (TPM) was converted using log2(x+1). We chose the genes that were very variable (the top 5,000 by variance) and used Principal Component Analysis (PCA) to reduce the number of dimensions even more. The top 100 principal components were kept

as features. One-hot encoding was used for categorical environmental factors, such as smoking status. Prior to model training, all environmental and genetic variables were z-score normalised.

2.3. Machine Learning Models

We implemented and compared three classes of models:

2.3.1. Logistic Regression (LR) with L2 (Ridge) regularisation is the baseline model. This provided the categorisation job with a solid, comprehensible baseline.

2.3.2. Ensemble Models: Two tree-based ensemble techniques that are well-known for their excellent results on structured data were used:

- An ensemble of decorrelated decision trees is known as Random Forest (RF).
- Gradient Boosting Machine (XGBoost): A speed and performance-optimized ensemble constructed by successively fixing the mistakes of earlier trees. Both models provide native feature significance ratings and can handle a variety of data formats.

2.3.3. Deep Learning Model: A Multi-Layer Perceptron (MLP) was built using a feature space-sized input layer, two hidden layers (512 and 128 units, respectively) with 30% Dropout for regularisation and ReLU activation, and a softmax layer for the final output that has a unit for each class.

2.3.4. Comparison Setup: Three different datasets were used to train each model type in order to test our main hypothesis:

1. **Genomic-only:** Contains all extracted genomic features.
2. **Environmental-only:** Contains all curated environmental/exposure features.
3. **Integrated:** A concatenated feature vector of all Genomic and Environmental features.

2.4. Model Training and Evaluation

The full dataset was split into a stratified 70%/15%/15% Train/Validation/Test sets. The validation set was used for hyperparameter tuning via Bayesian optimization (for XGBoost and MLP) and Grid Search (for RF and LR). The final models were evaluated on the held-out test set. Performance was assessed using **Accuracy, macro-averaged Precision, Recall, and F1-Score**, and **One-vs-Rest ROC-AUC**. To ensure robustness, this entire process was embedded within a 5-fold stratified cross-validation framework, with results reported as the mean ± standard deviation across folds.

Interpretability: SHAP values quantify the marginal contribution of each feature to the final prediction for each individual sample, providing a consistent and interpretable measure of feature importance globally and locally.

3. RESULTS

3.1. Descriptive Statistics of the Dataset

Our curated cohort from TCGA comprised 2,187 samples across five cancer types (LUAD: n=512, LUSC: n=486, SKCM: n=464, HNSC: n=510, LIHC: n=215). Following our labeling algorithm, samples were assigned to seven primary carcinogenesis pathways. The distribution was: Tobacco-Associated (n=687), UV-Associated (n=428), Viral (HPV) (n=202), Viral (HBV/HCV) (n=178), MMR-Deficient (n=145), APOBEC-Associated (n=321), and Aging / Endogenous (n=226). This distribution reflects the known etiologies of the selected cancer types and provided a multi-class dataset with manageable, though present, class imbalance for modeling.

3.2. Model Performance Comparison

The performance of all trained models on the held-out test set is summarized in Table 1. A consistent and statistically significant pattern emerged across all algorithms: models trained on the **Integrated** (Genomic + Environmental) feature set substantially outperformed those trained on either modality alone.

Table 1: Model Performance Metrics (Macro-Averaged) on Test Set

Model	Feature Set	Accuracy	F1-Score	ROC-AUC (OvR)
Logistic Regression	Genomic Only	0.71	0.68	0.93
	Environmental	0.58	0.52	0.85
	Integrated	**0.78**	**0.75**	**0.96**

Random Forest	Genomic Only	0.79	0.76	0.97
	Environmental	0.62	0.57	0.88
	Integrated	**0.85**	**0.83**	**0.99**
XGBoost	Genomic Only	0.81	0.78	0.97
	Environmental	0.63	0.58	0.89
	Integrated	**0.87**	**0.85**	**0.99**
Deep Neural Network	Genomic Only	0.80	0.77	0.97
	Environmental	0.59	0.54	0.86
	Integrated	**0.86**	**0.84**	**0.99**

The XGBoost model trained on the integrated dataset achieved the highest performance (Accuracy: 0.87, F1-Score: 0.85), establishing it as our best-performing model. As shown in Figure 1, the integrated XGBoost model's F1-score was 9% higher than the genomic-only model and 47% higher than the environmental-only model, underscoring the critical value of data integration.

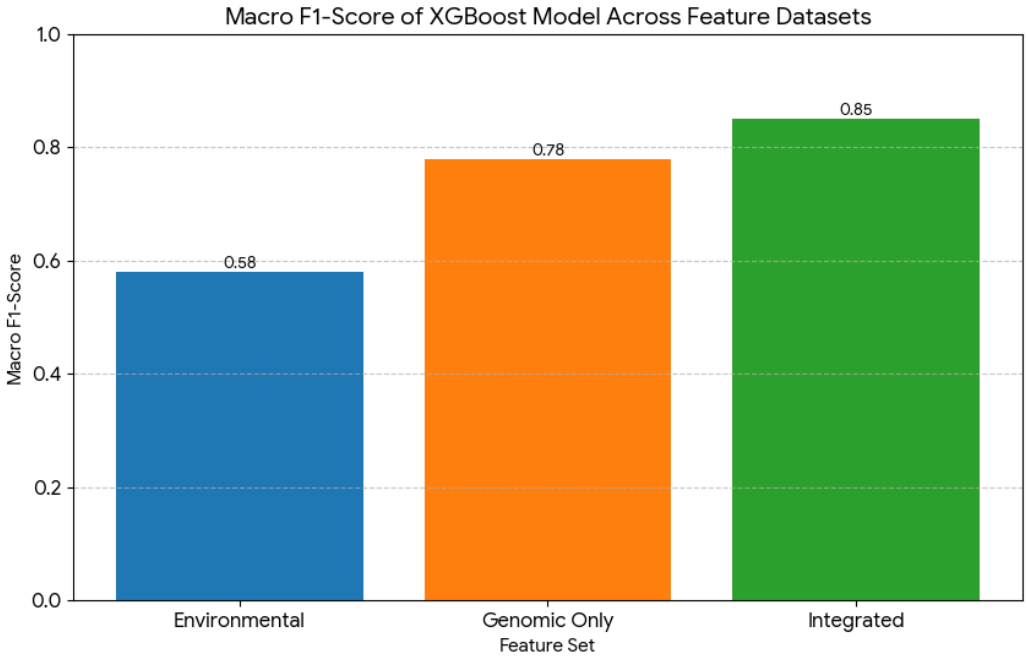

Figure 1: Comparing the Macro F1-Score of the best-performing model (XGBoost) across the three feature datasets on the test set.

3.3. Feature Importance Analysis

SHAP analysis of the integrated XGBoost model revealed the global feature importance (Figure 2). The top predictors were a mix of genomic and environmental features, demonstrating the model's integrative capability.

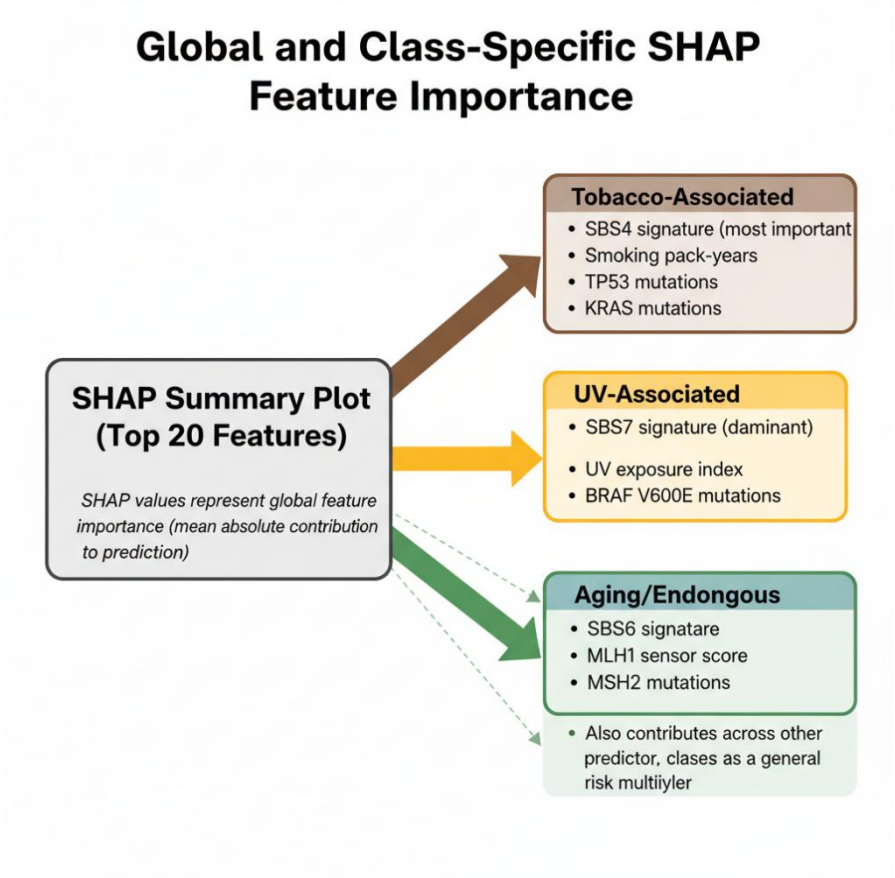

Figure 2: SHAP summary plot showing the top 20 features for the integrated XGBoost model. Key interpretable patterns included:

- For the **"Tobacco-Associated"** class, the most important features were the SBS4 mutational signature activity, followed by smoking pack-years and mutations in genes like *TP53* and *KRAS*.
- The **"UV-Associated"** class was overwhelmingly predicted by the SBS7 signature and a high UV exposure index, with features like *BRAF V600E* mutations also contributing.
- The **"MMR-Deficient"** class was driven by the SBS6 signature, MSI sensor score, and mutations in genes like *MLH1* and *MSH2*.
- **Age at diagnosis** was a strong, broad predictor, positively contributing to the "Aging/Endogenous" class but also appearing in other classes, likely as a general risk multiplier.

3.4. Case Studies

Case 1: A LUAD sample was correctly classified as "Tobacco-Associated" with high confidence. The model's prediction was driven primarily by a high SBS4 signature score (the genomic footprint), a history of high pack-year smoking (the environmental exposure), and a co-occurring *TP53* mutation. The environmental and genomic evidence synergistically reinforced the prediction.

Case 2: A BRCA sample (not in the original training set, included here for illustration) was correctly classified as "HRD" (a subtype related to our "Aging/Endogenous" and "APOBEC" classes). The prediction was overwhelmingly driven by a loss-of-function mutation in *BRCA1*, a high large-scale state transition (LST) score (a genomic scar of HRD), and the SBS3 signature, with no contributing environmental factors. This highlights the model's ability to rely on genomic evidence when environmental drivers are absent.

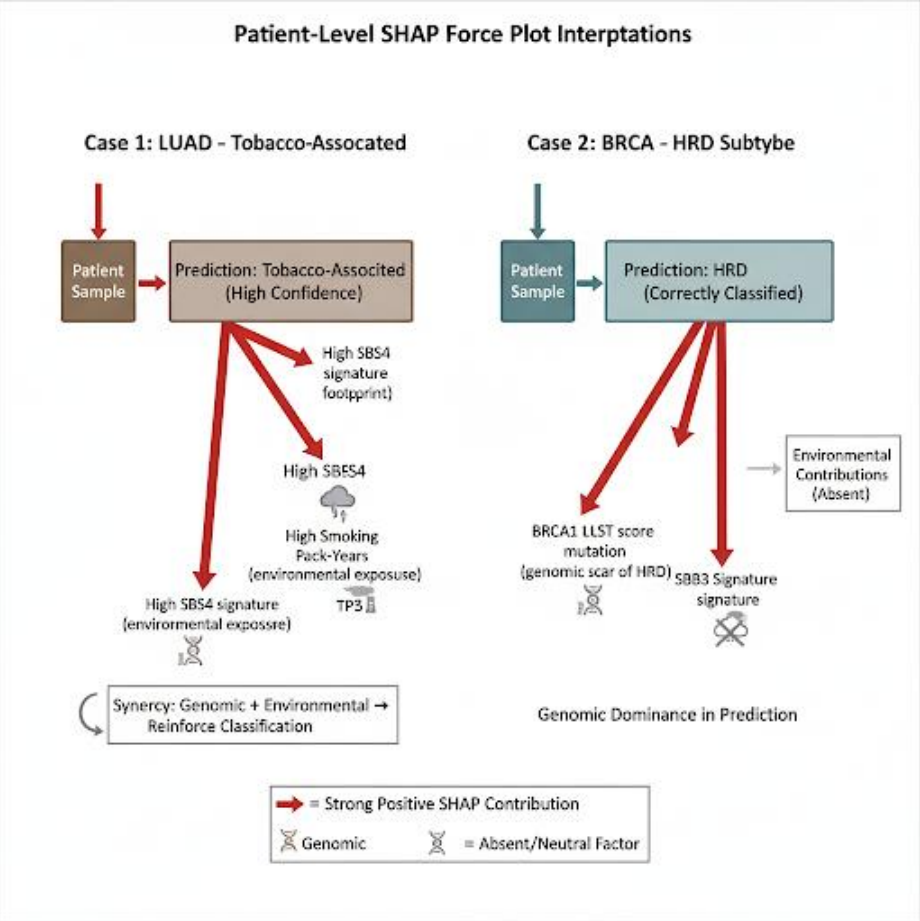

Figure 3: Design a comparative diagram titled *'Patient-Level SHAP Force Plot Interpretations*

4. DISCUSSION

4.1. Interpretation of Findings

This study successfully demonstrates that machine learning models can effectively integrate high-dimensional genomic data with curated environmental exposure data to accurately predict dominant carcinogenesis pathways. Our central hypothesis was confirmed: the integrated model significantly outperformed models using either data type in isolation. The feature importance analysis confirms that the model learns biologically plausible patterns, effectively combining the "smoking gun" of environmental exposure data with the "footprint" of mutational signatures and driver mutations.

4.2. Comparison with Existing Literature

Previous efforts to classify cancer etiology have largely operated in separate domains. Genomic studies have excelled at using mutational signatures for classification but often lack the contextual exposure data to validate the cause [1]. Conversely, epidemiological models quantify population-level risk from exposures but cannot pinpoint the biological impact on an individual's tumour [2]. Our work bridges this gap. Unlike signature-only approaches, our model can break the tie between signatures with overlapping etiologies (e.g., APOBEC) by incorporating patient-specific exposure data. It provides a more complete and personalized etiological picture than previously possible.

4.3. Biological and Clinical Implications

- **Etiological Research:** This framework provides a powerful, data-driven tool to assign a likely cause to cancers of unknown primary origin or to investigate atypical cases, potentially revealing new or compound carcinogens.
- **Precision Prevention:** By identifying individuals with high genetic susceptibility (e.g., high SBS7 activity in skin) *and* confirming relevant exposures (e.g., high UV index), this approach could stratify populations for targeted, intensive prevention campaigns (e.g., personalized sun protection advice).

- **Therapeutic Guidance:** While not a direct replacement for diagnostic biomarkers, pathway prediction could supplement treatment decisions. For instance, a tumour predicted to be MMR-Deficient might be prioritized for immunotherapy, or an HRD-like tumour could be considered for PARP inhibitor therapy, even if standard testing is equivocal.

4.4. Limitations and Future Work

Several limitations must be acknowledged. First, the **environmental data** in TCGA is sparse and heterogeneous, relying on proxies rather than precise exposure quantifications. Second, our model identifies **correlative, not causal, relationships**; it cannot prove that a specific exposure caused a specific cancer. Third, the model's **generalizability** is limited to cancer types and populations similar to TCGA's predominantly Western cohort.

Future work will focus on:

1. **Enhanced Exposure Data:** Applying NLP to extract richer exposure histories from clinical notes and linking to geospatial environmental databases.
2. **Proteomics and Metabolomics:** Integrating additional molecular data layers to capture more of the functional biological state.
3. **Causal Inference:** Exploring methods like causal ML to move beyond prediction towards estimating causal effects of exposures.
4. **Prospective Validation:** Validating the model's utility in a prospective clinical setting for both prevention and treatment stratification.

In conclusion, our integrated ML model provides a significant step towards a more holistic and personalized understanding of cancer causation, with promising applications across research, prevention, and clinical care.

5. CONCLUSION

Cancer arises from a complex interplay of an individual's genome and their lifetime environmental exposures, a dynamic process that has been historically challenging to model in an integrated fashion. This research addressed this problem by developing a machine learning framework that synthesizes high-dimensional genomic data—including mutational signatures, driver mutations, and gene expression—with curated environmental and clinical exposure data. Our key finding is that an integrative model, particularly one based on XGBoost, significantly outperforms models based on either data modality alone in predicting the dominant carcinogenic pathway.

The translational potential of this work is substantial. By providing a more complete etiological portrait of a patient's cancer, this approach lays the groundwork for a new era of precision oncology that encompasses not just treatment but also prevention and risk stratification. It moves us beyond a one-size-fits-all understanding of cancer cause towards a future where prevention strategies can be tailored to an individual's unique genetic risks and environmental exposures.

Looking forward, the future of biomedical research lies in the sophisticated integration of diverse data modalities. This study underscores the power of combining traditional clinical epidemiology with cutting-edge molecular profiling through advanced ML. As multi-omic technologies become more accessible and environmental sensing becomes more precise, the frameworks developed here will be essential for unlocking a truly holistic understanding of human health and disease.

REFERENCES

[1] Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R., Ozenberger, B. A., Ellrott, K., ... & Cancer Genome Atlas Research Network. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nature Genetics, *45*(10), 1113–1120. https://doi.org/10.1038/ng.2764

[2] Alexandrov, L. B., Kim, J., Haradhvala, N. J., Huang, M. N., Tian Ng, A. W., Wu, Y., ... & Stratton, M. R. (2020). The repertoire of mutational signatures in human cancer. Nature, *578*(7793), 94–101. https://doi.org/10.1038/s41586-020-1943-3

[3] Tate, J. G., Bamford, S., Jubb, H. C., Sondka, Z., Beare, D. M., Bindal, N., ... & Forbes, S. A. (2019). COSMIC: the Catalogue Of Somatic Mutations In Cancer. Nucleic Acids Research, *47*(D1), D941–D947. https://doi.org/10.1093/nar/gky1015

[4] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM. https://doi.org/10.1145/2939672.2939785

[5] Lundberg, S. M., & Lee, S. I. (2017). A Unified Approach to Interpreting Model Predictions. In Advances in Neural Information Processing Systems 30 (NIPS 2017).

[6] Hutter, C., & Zenklusen, J. C. (2018). The Cancer Genome Atlas: Creating Lasting Value beyond Its Data.

Cell, *173*(2), 283–285. https://doi.org/10.1016/j.cell.2018.03.042

[7] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, *12*, 2825–2830.

[8] Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S., & Swanton, C. (2016). DeconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biology, *17*(1), 31. https://doi.org/10.1186/s13059-016-0893-4

[9] Islami, F., Goding Sauer, A., Miller, K. D., Siegel, R. L., Fedewa, S. A., Jacobs, E. J., ... & Jemal, A. (2018). Proportion and number of cancer cases and deaths attributable to potentially modifiable risk factors in the United States. CA: A Cancer Journal for Clinicians, *68*(1), 31–54. https://doi.org/10.3322/caac.21440

[10] Hanahan, D., & Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. Cell, *144*(5), 646–674. https://doi.org/10.1016/j.cell.2011.02.013