



OPEN

# Multimodal artificial intelligence-based pathogenomics improves survival prediction in oral squamous cell carcinoma

Andreas Vollmer<sup>1✉</sup>, Stefan Hartmann<sup>1</sup>, Michael Vollmer<sup>2</sup>, Veronika Shavlokhova<sup>3</sup>, Roman C. Brands<sup>1</sup>, Alexander Kübler<sup>1</sup>, Jakob Wollborn<sup>4</sup>, Frank Hassel<sup>5</sup>, Sebastien Couillard-Despres<sup>6,7</sup>, Gernot Lang<sup>8</sup> & Babak Saravi<sup>4,5,6,8</sup>

In this study, we aimed to develop a novel prognostic algorithm for oral squamous cell carcinoma (OSCC) using a combination of pathogenomics and AI-based techniques. We collected comprehensive clinical, genomic, and pathology data from a cohort of OSCC patients in the TCGA dataset and used machine learning and deep learning algorithms to identify relevant features that are predictive of survival outcomes. Our analyses included 406 OSCC patients. Initial analyses involved gene expression analyses, principal component analyses, gene enrichment analyses, and feature importance analyses. These insights were foundational for subsequent model development. Furthermore, we applied five machine learning/deep learning algorithms (Random Survival Forest, Gradient Boosting Survival Analysis, Cox PH, Fast Survival SVM, and DeepSurv) for survival prediction. Our initial analyses revealed relevant gene expression variations and biological pathways, laying the groundwork for robust feature selection in model building. The results showed that the multimodal model outperformed the unimodal models across all methods, with c-index values of 0.722 for RSF, 0.633 for GBSA, 0.625 for FastSVM, 0.633 for CoxPH, and 0.515 for DeepSurv. When considering only important features, the multimodal model continued to outperform the unimodal models, with c-index values of 0.834 for RSF, 0.747 for GBSA, 0.718 for FastSVM, 0.742 for CoxPH, and 0.635 for DeepSurv. Our results demonstrate the potential of pathogenomics and AI-based techniques in improving the accuracy of prognostic prediction in OSCC, which may ultimately aid in the development of personalized treatment strategies for patients with this devastating disease.

**Keywords** Oral cancer, Oral squamous cell carcinoma, Survival, Artificial intelligence, Deep learning, Machine learning, Pathogenomics, Multimodal prediction

## Abbreviations

OSCC	Oral squamous cell carcinoma
AI	Artificial intelligence
TCGA	The cancer genome atlas
H&E	Hematoxylin and eosin
HPV	Human papillomavirus
RSF	Random survival forest
GBSA	Gradient boosting survival analysis

<sup>1</sup>Department of Oral and Maxillofacial Plastic Surgery, University Hospital of Würzburg, 97070 Würzburg, Franconia, Germany. <sup>2</sup>Department of Oral and Maxillofacial Surgery, Tuebingen University Hospital, Osianderstrasse 2-8, 72076 Tuebingen, Germany. <sup>3</sup>Maxillofacial Surgery University Hospital Ruppiner-Brandenburg, Fehrbelliner Straße 38, 16816 Neuruppin, Germany. <sup>4</sup>Department of Anesthesiology, Perioperative and Pain Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, USA. <sup>5</sup>Department of Spine Surgery, Loretto Hospital, Freiburg, Germany. <sup>6</sup>Institute of Experimental Neuroregeneration, Paracelsus Medical University, 5020 Salzburg, Austria. <sup>7</sup>Austrian Cluster for Tissue Regeneration, Vienna, Austria. <sup>8</sup>Department of Orthopedics and Trauma Surgery, Medical Center - University of Freiburg, Faculty of Medicine, University of Freiburg, Freiburg, Germany. ✉email: Vollmer\_a@ukw.de

Cox PH	Cox proportional hazards
FastSVM	Fast survival support vector machine
DeepSurv	Deep survival analysis
PCA	Principal component analysis
TCGAbiolinks	The cancer genome atlas bioinformatics links
GO	Gene ontology
KEGG	Kyoto encyclopedia of genes and genomes
DAVID	Database for annotation, visualization, and integrated discovery
SPSS	Statistical package for the social sciences
N	Number
CI	Confidence interval
AJCC	American joint committee on cancer
DE	Differentially expressed
GOTERM_MF_DIRECT	Gene ontology molecular function direct
GOTERM_CC_DIRECT	Gene ontology cellular component direct
UP_KW_CELLULAR_COMPONENT	UniProt keyword cellular component
KEGG_PATHWAY	Kyoto encyclopedia of genes and genomes pathway
GOTERM_BP_DIRECT	Gene ontology biological process direct
UP_KW_MOLECULAR_FUNCTION	UniProt keyword molecular function
UP_SEQ_FEATURE	UniProt sequence feature
UP_KW_DOMAIN	UniProt keyword domain
CPH	Cox proportional hazards
CoxPHSurvivalAnalysis	Implementation of cox proportional hazards model by CoxPHSurvivalAnalysis from sksurv.linear_model
sksurv	Scikit-survival
RandomSurvivalForest	Random survival forest model
GradientBoostingSurvivalAnalysis	Gradient boosting survival analysis model
FastSurvivalSVM	Fast survival support vector machine model
KerasRegressor	Keras regressor model
SCC	Squamous cell carcinoma
HFBSurv	Hierarchical factorized bilinear fusion for cancer survival prediction

Oral cancer is a significant global health concern with high morbidity and mortality rates. Oral squamous cell carcinoma (OSCC), the most common type of oral cancer, results in an estimated 378,000 new diagnoses and over 177,000 deaths worldwide annually<sup>1</sup>. OSCC is commonly associated with unhealthy habits such as alcohol abuse, tobacco use, and chewing betel nuts, as well as human papillomavirus (HPV) infection<sup>2</sup>. The development of OSCC is generally asymptomatic in the early stages, leading to late diagnosis, extensive lesions, and potential metastases<sup>3</sup>. Despite intervention with advanced treatment regimens, the survival rate of OSCC has not significantly improved in recent decades, underlining the limitations of current prognostic methods<sup>4</sup>. These traditional approaches, primarily based on clinicopathological factors such as demographic variables, tumor size, lymph node involvement, and metastasis, often fail to capture the complex biological heterogeneity of OSCC, leading to suboptimal treatment stratification and prognostication<sup>5</sup>.

Prognostic markers are urgently required to better adjust treatment intensity and avoid serious complications caused by overtreatment. The current gold standard for cancer diagnosis involves the manual examination of H&E-stained slides by pathologists<sup>6</sup>. However, recent advances in deep learning for digital pathology have allowed for the use of whole-slide images (WSIs) for computational image analysis tasks, such as cellular segmentation and tissue classification<sup>7</sup>. Genomic data can provide a deeper molecular characterization of the tumor, offering the potential for prognostic and predictive biomarker discovery.

The utilization of unimodal input, which involves relying on data from a single resource, fails to fully exploit the potential benefits of incorporating more extensive information from other aspects of patients that may impact their overall survival time<sup>8</sup>. Current survival prediction in oncology often relies on traditional methods like the Kaplan–Meier estimator or Cox proportional hazards (Cox-PH) model. While these approaches have been the cornerstone of cancer prognosis, they primarily depend on limited variables such as patient demographics, tumor stage, and histopathological risk factors. This conventional methodology lacks the capability to account for the vast heterogeneity and complex biological mechanisms underlying different cancer types, including OSCC<sup>9</sup>. Recent research findings suggest that leveraging multi-omics data of cancer can significantly enhance the accuracy of non-small-cell lung cancer subtype classification compared to using a single modality approach<sup>10</sup>. Multimodal survival prediction is a sophisticated method used for biomarker discovery, patient stratification, and therapeutic response prediction<sup>9</sup>. Artificial intelligence-processed pathogenomics is a relatively new research field that combines genomics and pathology and has shown promise in identifying novel biomarkers and therapeutic targets for cancer<sup>11</sup>. Recent studies have demonstrated the potential of pathogenomics in predicting the survival of patients with different cancer types<sup>11</sup>. The increasing availability of high-throughput, multidimensional data from initiatives like the cancer genome atlas (TCGA) has revolutionized the field of cancer research. However, the complexity and volume of such data exceed the capabilities of traditional survival analysis methods. This gap has paved the way for the integration of advanced AI techniques, especially deep learning (DL), in analyzing complex clinical and genomic data. Models combining neural network architectures with the Cox-PH model, such as DeepSurv and Cox-nnet, have shown promise in outperforming traditional models by leveraging more complex, non-linear relationships in the data<sup>12</sup>. These developments underscore the potential

of a multimodal approach, integrating clinical characteristics with diverse omics data, for enhancing cancer prognosis predictions. Particularly for OSCC, where traditional prognostic models have limitations, leveraging artificial intelligence (AI)-processed pathogenomics—an innovative field that combines genomics and pathology—holds great promise. This approach, relatively unexplored in OSCC, has shown potential in other cancer types for improving survival prediction accuracy<sup>9,12</sup>. Thus, utilizing a multimodal data integration strategy, which includes clinical data, histology, and genetic information, can potentially overcome the limitations of current prognostic models and pave the way for more precise, personalized treatment strategies, ultimately leading to improved patient outcomes.

AI-based techniques, such as machine learning, have been increasingly applied to various fields of medicine, including cancer research, to enhance the accuracy of diagnosis, treatment selection, and prognosis prediction<sup>13</sup>. Utilizing multimodal data as input for AI-based algorithms could be a novel and groundbreaking approach for survival prediction. However, few methods have been proposed to fully exploit the potential of multiple data modalities<sup>8</sup>.

The primary objective of our study is to enhance the prognostic prediction in OSCC by leveraging multimodal data encompassing clinical, histological, and genetic information. To achieve this, we first undertook a thorough exploration of gene expression profiles and biological processes in OSCC. This initial phase involves comprehensive gene expression analyses, principal component analyses, gene enrichment studies and feature selection. These steps are pivotal in identifying key genetic features that might underpin OSCC pathogenesis, offering critical insights into the disease's complexity. Subsequently, we employ these insights to inform our machine learning and deep learning models. By first establishing a deep understanding of the underlying genetic and histopathological landscape, our approach aims to refine the selection of features that are most indicative of survival outcomes. This methodical progression from fundamental gene expression studies to the application of advanced AI techniques is designed to ensure that the resulting models are not only technically robust but also grounded in clinically relevant biological insights. The results of this study have the potential to provide novel insights into the development of prognostic and predictive biomarkers for OSCC, which can aid in the development of more personalized treatment plans and improve patient outcomes.

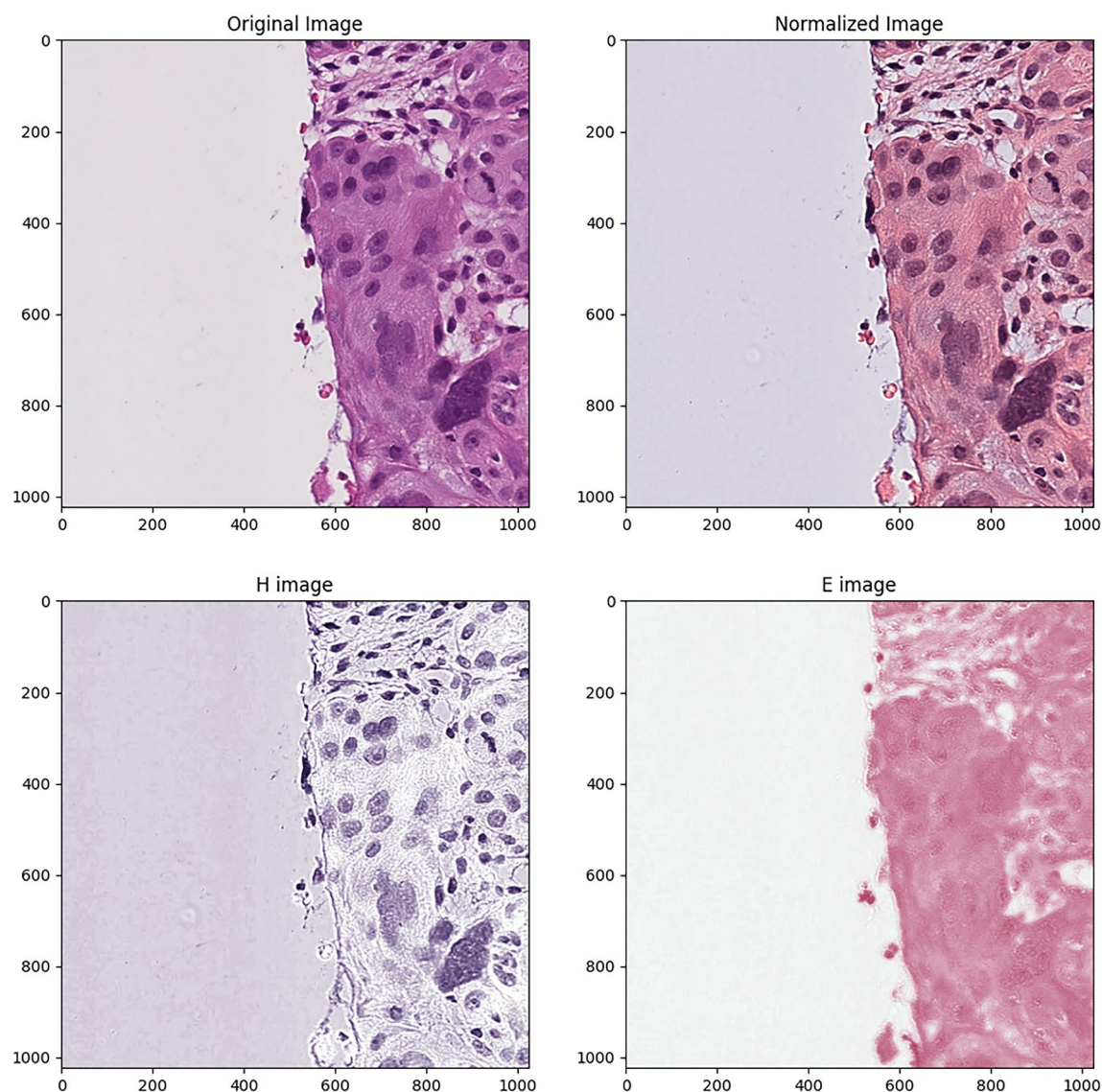
## Methods

### Study design

The original datasets comparing the gene expression profiles between solid, healthy, and solid tumor tissue were obtained from the National Cancer Institute GDC Data Portal (<https://portal.gdc.cancer.gov/>). All data that were processed were from the TCGA-HNSC project, which included only head and neck squamous cell carcinomas<sup>14</sup>. TCGA utilizes a strict set of criteria for inclusion into the study due to the rigorous and comprehensive nature of the work being performed. Tissue samples from tumors and their corresponding germline DNA sources are collected and handled by the Centralized Biorepository, a dedicated facility responsible for examining specimen information and processing all samples to maintain uniform pathology evaluation and production of molecular elements (DNA and RNA). Upon arrival at the Centralized Biorepository, every sample undergoes a stringent quality assurance process before being approved for comprehensive analysis within the TCGA workflow. A pathologist examines each specimen to verify the diagnosis and ensure it fulfills the inclusion criteria. Specifically, TCGA mandates that samples possess a minimum of 60% tumor nuclei and no more than 20% necrotic tissue. Once a sample clears the pathological assessment, nucleic acids are extracted, and genotyping is carried out to accurately link each tumor specimen with its corresponding normal tissue. An important goal in establishing this central resource is to ensure that molecular analytes (i.e., DNA and RNA) extracted from tissue samples are of consistent and high quality. Next, these analytes undergo a molecular quality control process and then are distributed to TCGA Cancer Genome Characterization Centers and Genome Sequencing Centers for genomic analysis. All samples in TCGA have been collected and utilized following strict policies and guidelines for the protection of human subjects, informed consent, and IRB review of protocols<sup>14</sup>. Inclusion criteria for the present study following the extraction of the initial TCGA-HNSC dataset were OSCC and patients who had histopathology and genetics data available. In alignment with previous prognostic research on OSCC utilizing the TCGA-HNSC dataset<sup>15,16</sup>, only the following sites were included: alveolar ridge, base of tongue, buccal mucosa, floor of mouth, hard palate, hypopharynx, lip, oral cavity, oral tongue, oropharynx, and tonsil. No additional exclusion criteria were applied beyond these parameters, allowing for a comprehensive and representative sample of OSCC patients for our analyses.

### Image processing

Digitized whole-slide images of H&E-stained specimens from primary untreated tumors were processed to extract quantitative histological features, sourced from the TCGA database. Employing a custom Python script and the OpenSlide library, we applied color normalization techniques to these images, following the methodologies described by Macenko et al. and implemented in Python by Vahadane et al.<sup>17,18</sup>. This process ensured consistent color representation across slides. To facilitate feature analysis, images were segmented into tiles of 1024 by 1024 pixels, focusing on areas with the highest density of diagnostic information as identified in previous research<sup>19</sup>. Figure 1 illustrates the normalization of a random tile by the algorithm. Using CellProfiler<sup>20</sup>, we extracted 170 quantitative features from these tiles, including metrics related to cell shape, size, texture, and pixel intensity distributions. This multi-dimensional data was then integrated with genomic and clinical information for comprehensive analysis. Detailed methodologies and scripts used for image processing are available in the Supplementary Material.



**Figure 1.** Representative examples of a non-normalized tile, a normalized tile, the Hematoxylin (H)-stained tile, and the Eosin (E)-stained tile. The non-normalized tile represents the original raw image tile extracted from the digital whole slide image. The H tile is generated by first converting the original histology tile to grayscale and then applying a high-pass filter, which enhances the high-frequency information in the image. This results in an image with a blue-purple hue, as hematoxylin stains the nuclei of cells in shades of blue. The H tile emphasizes the cell nuclei, which contain important diagnostic information. The E tile, on the other hand, is generated by first converting the original histology tile to grayscale and then applying a low-pass filter, which retains the low-frequency information in the image. This results in an image with a pink-orange hue, as eosin stains the cytoplasm and extracellular matrix in shades of pink. The E tile emphasizes the tissue structure and texture, which can provide additional diagnostic information. The normalized tile is the result of the normalization of the tile to reduce color variation between slides and was used for further analyses. The figure was generated using Python (version 3.10.4).

## Genomics analyses

Our genomic analysis utilized RNA-Seq data from the TCGA-HNSC project, focusing on primary tumor and normal tissue samples. Data preprocessing and analysis were conducted using the TCGAtoolkit package in R, employing a series of steps to ensure data quality and relevance. Lowly expressed genes were filtered out using the filterByExpr method in the limma package to concentrate on genes with significant expression levels. The TMM method followed by the voom transformation was applied for normalization, adjusting for library compositional differences and preparing data for linear modeling. Employing linear modeling and empirical Bayes statistics, we identified the top 200 differentially expressed genes. These genes were further analyzed through PCA to visualize variance and clustering, aiding in distinguishing between tumor and healthy tissue samples. To assess gene expression's impact on survival, we applied the Elastic Net algorithm<sup>21</sup>. Elastic Net, with its dual advantages of Lasso's feature selection and Ridge's multicollinearity management, provides a balanced approach that enhanced

both the interpretability and robustness for predictions<sup>22</sup>, making it especially suited for the complex nature of OSCC gene expression data. The analysis led to the identification of 72 predictive genes, visualized through a heatmap created with the heatmap.2 function from the gplots package, highlighting the expression patterns between normal and tumor samples.

Gene enrichment analysis was conducted using the DAVID (Database for Annotation, Visualization, and Integrated Discovery) bioinformatics database to identify significant GO terms and KEGG pathways<sup>23–25</sup> among the differentially expressed genes, setting a significance threshold at  $p < 0.05$ . This comprehensive genomic analysis approach, detailed further in the Supplementary Material, allowed for the robust identification and visualization of key genes and pathways relevant to OSCC.

### Statistical analyses and artificial intelligence-based techniques

Our analysis utilized a combination of statistical methods and artificial intelligence-based techniques, executed in R (version 3.2.3), Python (version 3.10.4), and SPSS Modeler. Supported by high-performance computing, including an AMD Ryzen 9 5950X processor and NVIDIA GeForce RTX 3090 GPU, we processed and analyzed OSCC data for predictive modeling and survival analysis. Data preprocessing, involving cleaning and normalization, was conducted using scikit-learn and Pandas libraries. We employed survival prediction models such as Random Survival Forest, Gradient Boosting Survival Analysis, Survival Support Vector Machine, Cox proportional hazards model, and a custom-developed deep learning model in Keras, focusing on the Cox model's negative log partial likelihood for patient outcome prediction. Model performance evaluation was based on the concordance index (C-index), with feature importance assessed through a c-index reduction approach to refine model predictions. We utilized a comprehensive strategy to address model overfitting and selection bias, incorporating regularization techniques, manual hyperparameter tuning, and k-fold cross-validation. This analytical framework facilitated the integration of clinical, histological, and genetic data into our models. For a detailed description of the data preprocessing steps, model development, and evaluation criteria, refer to the Supplementary Material.

## Results

### Descriptive statistics

Table 1 illustrates the descriptive statistics of the analyzed TCGA dataset. A total of 406 OSCC patients were analyzed.  $N = 294$  (72.41%) were male, and  $n = 112$  (27.59%) were female. The mean age of patients at the time of diagnosis was  $61.53 \pm 12.38$  years. The majority of patients were classified as "white race" ( $n = 354$ ; 87.19%), followed by "black or African American" ( $n = 29$ ; 7.14%). The most frequent pathological ( $n = 196$ ; 48.28%) and clinical ( $n = 203$ ; 50.00%) stage was IV A.  $N = 17$  (4.19%) patients had prior malignancies, and  $n = 8$  (1.97%) received prior treatment.  $N = 139$  (34.24%) had no signs of pathological lymph node metastases, and  $n = 145$  (35.71%) were classified as M0 based on pathological AJCC staging. Figure 2 shows the Kaplan–Meier survival curve and the risk table of the cohort. Time: in days. The median Survival estimate according to the Kaplan–Meier-Method was 1591 days (95% CI 1199.89–1982.11).

### Comprehensive analysis of gene expression profiles: identifying key differentially expressed genes and pathways in OSCC

Figure 3 highlights the results of the PCA. The two average circles in the PCA analysis represent the centroids of the two groups. They show the average position of the data points in each group along the first two principal components. The centroid is calculated by taking the mean of the x and y coordinates of all the data points in the group. As the circles are far apart, it suggests that the two groups are well separated along the first two principal components, which is a sign of differential gene expression between the two groups. Figure 4 highlights the top differentially expressed genes that were obtained through the comparison of solid normal and tumor tissue by the ElasticNet model. The further analyses contained a total of 200 differentially expressed genes that were assessed solely for the tumor tissue samples.

The results of the gene enrichment analyses are shown in Fig. 5. The results showed that several biological processes and molecular functions were significantly enriched. The most enriched molecular function was protein binding, which was identified in 65.2% of the analyzed genes. The cellular component analysis revealed that the plasma membrane, secreted proteins, and extracellular regions were highly represented. Interestingly, metabolic pathways were also enriched, suggesting a possible link between metabolic processes and OSCC development that was also suggested recently through genomics analyses<sup>26</sup>. In addition, lipid metabolism, oxidoreductase activity, and cell junction were also found to be enriched.

### Evaluating prognostic factors and model performance in ai-based oscar survival prediction

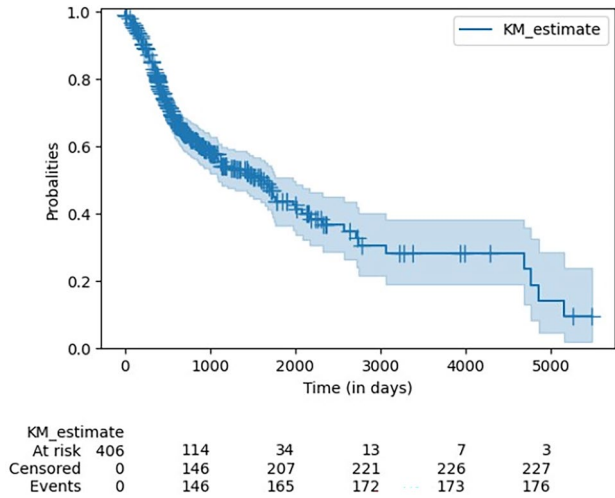
The results of the feature importance analyses for clinical features are shown in Fig. 6. As expected, the AJCC staging variables were the most significant predictors of survival. Furthermore, smoking and gender were among the top 10 predictors. This confirms prior knowledge that smoking, and gender are important predictors of survival<sup>27–29</sup>.

Table 2 shows the comparison of unimodal and multimodal artificial intelligence-based analyses for survival prediction. We assessed the performance of unimodal and multimodal models in predicting patient outcomes using the c-index metric. The unimodal models included clinical, pathology, or genetic features, while the multimodal model combined all three types of features. The results showed that the multimodal model outperformed the unimodal models across all methods, with c-index values of 0.722 for RSF, 0.633 for GBSA, 0.625 for FastSVM, 0.633 for CoxPH, and 0.515 for DeepSurv. When considering only important features, the multimodal model continued to outperform the unimodal models, with c-index values of 0.834 for RSF, 0.747 for GBSA, 0.718 for FastSVM, 0.742 for CoxPH, and 0.635 for DeepSurv. The important features in the multimodal model

Variable	Count (%)	Mean $\pm$ SD
Demographics		
Age at diagnosis		61.53 $\pm$ 12.38
Gender–Female	112 (27.59%)	
Gender–Male	294 (72.41%)	
Race–Black or African American	29 (7.14%)	
Race–White	354 (87.19%)	
Race–Asian	10 (2.46%)	
Race–American Indian or Alaska Native	1 (0.25%)	
Race–Missing	12 (2.96%)	
Ethnicity–Not Hispanic or Latino	360 (88.67%)	
Ethnicity–Hispanic or Latino	19 (4.68%)	
Ethnicity–Missing	27 (6.65%)	
Lifestyle factors and previous malignancy		
Cigarettes per day		1.30 $\pm$ 1.91
Years smoked		8.74 $\pm$ 15.86
Pack years smoked		23.71 $\pm$ 34.87
Alcohol history—No	130 (32.02%)	
Alcohol history—Yes	266 (65.52%)	
Alcohol history—Missing	10 (2.46%)	
Prior malignancy—No	389 (95.81%)	
Prior malignancy—Yes	17 (4.19%)	
Clinical staging		
AJCC clinical stage—Stage I	15 (3.69%)	
AJCC clinical stage—Stage II	88 (21.67%)	
AJCC clinical stage—Stage III	80 (19.70%)	
AJCC clinical stage—	203 (50.00%)	
AJCC clinical stage—Stage IVB	5 (1.23%)	
AJCC clinical stage—Stage IVC	4 (0.99%)	
AJCC clinical stage—Missing	11 (2.71%)	
AJCC clinical T–T1	29 (7.14%)	
AJCC clinical T–T2	130 (32.02%)	
AJCC clinical T–T3	97 (23.89%)	
AJCC clinical T–T4	22 (5.42%)	
AJCC clinical T–T4a	114 (28.08%)	
AJCC clinical T–T4b	2 (0.49%)	
AJCC clinical T–TX	9 (2.22%)	
AJCC clinical T–Missing	3 (0.74%)	
AJCC clinical N–N0	195 (48.03%)	
AJCC clinical N–N1	66 (16.26%)	
AJCC clinical N–N2	15 (3.69%)	
AJCC clinical N–N2a	13 (3.20%)	
AJCC clinical N–N2b	63 (15.52%)	
AJCC clinical N–N2c	31 (7.64%)	
AJCC clinical N–N3	4 (0.99%)	
AJCC clinical N–NX	16 (3.94%)	
AJCC clinical N–Missing	3 (0.74%)	
AJCC clinical M–M0	383 (94.33%)	
AJCC clinical M–M1	3 (0.74%)	
AJCC clinical M–MX	17 (4.19%)	
AJCC clinical M–Missing	3 (0.74%)	
Pathological staging		
AJCC pathologic stage–Stage I	23 (5.67%)	
AJCC pathologic stage–Stage II	61 (15.02%)	
AJCC pathologic stage–Stage III	70 (17.24%)	
AJCC pathologic stage–Stage IVA	196 (48.28%)	
AJCC pathologic stage–Stage IVB	8 (1.97%)	
Continued		

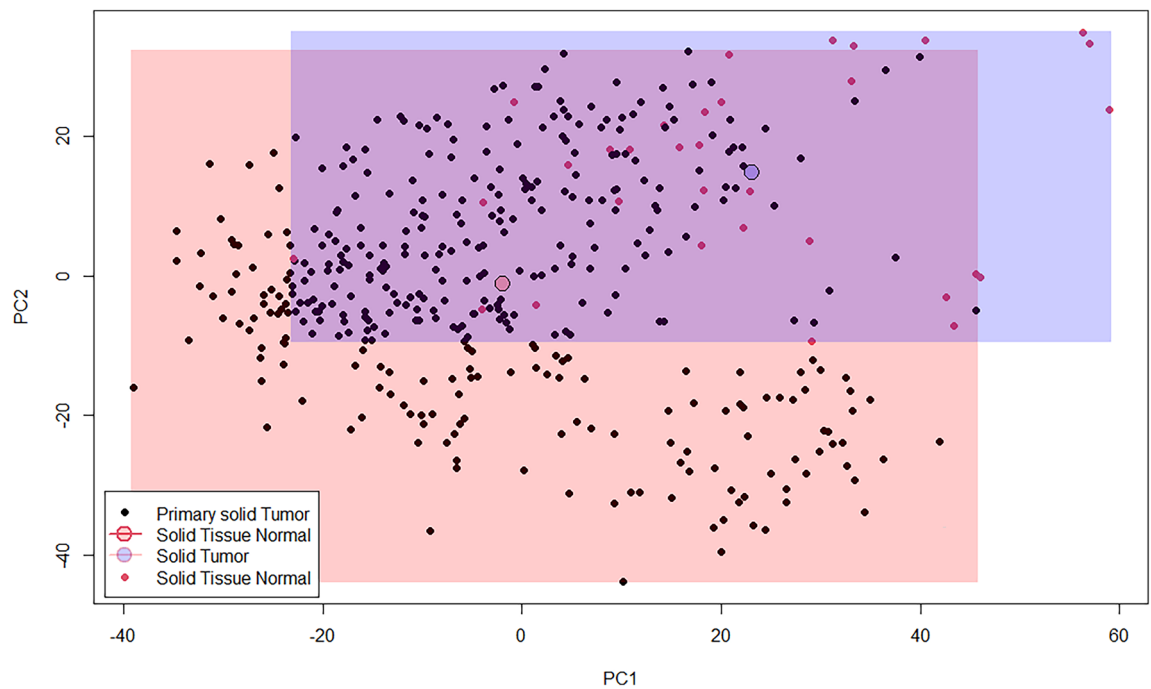
Variable	Count (%)	Mean ± SD
AJCC pathologic stage–Missing	48 (11.82)	
AJCC pathologic T–T1	38 (9.36%)	
AJCC pathologic T–T2	120 (29.56%)	
AJCC pathologic T–T3	79 (19.46%)	
AJCC pathologic T–T4	11 (2.71%)	
AJCC pathologic T–T4a	114 (28.08%)	
AJCC pathologic T–T4b	4 (0.99%)	
AJCC pathologic T–TX	21 (5.17%)	
AJCC pathologic T–Missing	19 (4.68%)	
AJCC pathologic N–N0	139 (34.24%)	
AJCC pathologic N–N1	56 (13.79%)	
AJCC pathologic N–N2	11 (2.71%)	
AJCC pathologic N–N2a	4 (0.99%)	
AJCC pathologic N–N2b	85 (20.94%)	
AJCC pathologic N–N2c	34 (8.37%)	
AJCC pathologic N–N3	4 (0.99%)	
AJCC pathologic N–NX	52 (12.81%)	
AJCC pathologic N–Missing	21 (5.17%)	
AJCC pathologic M–M0	145 (35.71%)	
AJCC pathologic M–MX	52 (12.81%)	
AJCC pathologic M–Not reported	209 (51.48%)	
Prior treatment		
Prior treatment–No	398 (98.03%)	
Prior treatment–Yes	8 (1.97%)	
Outcome		
Vital status–Dead	177 (43.60%)	
Vital status–Alive	229 (56.40%)	
Overall survival		878.87 ± 857.18

**Table 1.** Descriptive statistics of clinical features of patients (n = 406).



**Figure 2.** Kaplan–Meier survival curve of the cohort. The figure was generated using Python (version 3.10.4).

were ENSG00000150667.8 (fibrous sheath interacting protein 1), ENSG00000186868.16 (microtubule associated protein tau), ENSG00000119147.10 (ECRG4 augurin precursor), ENSG00000272540.1 (novel transcript antisense to TUBB), ENSG00000105929.16 (ATPase H + transporting V0 subunit a4), ENSG00000124203.6 (zinc finger protein 831), ENSG00000172340.15 (succinate-CoA ligase GDP-forming subunit beta), Intensity MinIntensity Eosin (minimum pixel intensity values for Eosin staining), years of smoking, and AJCC clinical



**Figure 3.** Principal component analysis (PCA) of differentially expressed genes. Each principal component (PC) represents a linear combination of the original variables (gene expression levels) and is orthogonal to the other components. PC1 and PC2 are the two linear combinations of the gene expression data that explain the most variation in the dataset. The axes in a PCA plot represent the principal components. The x-axis represents the first principal component (PC1) and the y-axis represents the second principal component (PC2). Each point in the plot corresponds to a sample, and its position along the axes represents its scores on the principal components. Points that are close together on the plot have similar gene expression profiles, while points that are further apart have more distinct profiles. The rectangles represent the boundaries of each group along the two principal components. The two average circles in the PCA analysis represent the centroids of the two groups. They show the average position of the data points in each group along the first two principal components. The figure was generated using Python (version 3.10.4).

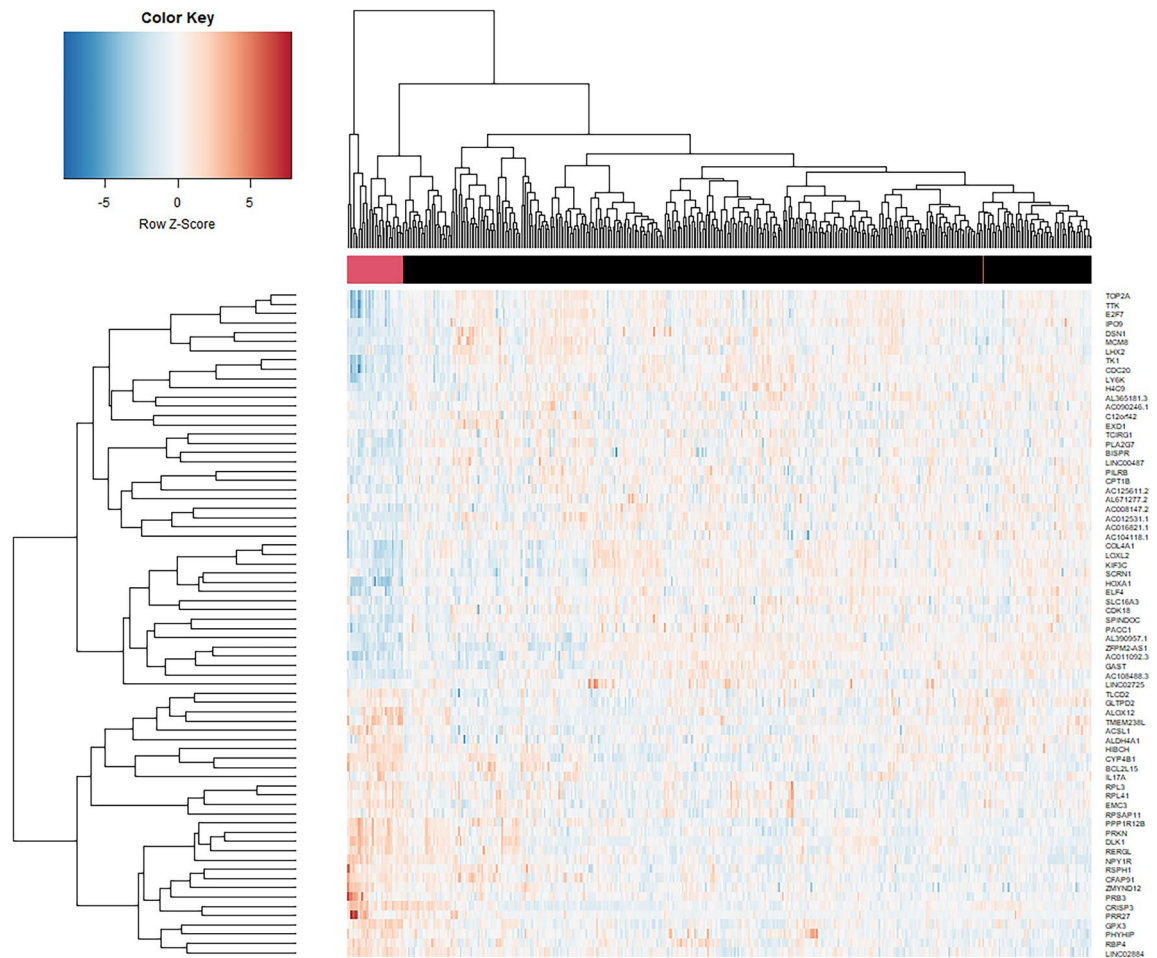
N-staging. These results suggest that combining clinical, pathology, and genetic features improves the accuracy of predicting patient outcomes compared to using each feature type alone.

Figure 7 illustrates the pooled multimodal feature importance as evaluated by the models. The heatmap displays the pooled feature importance scores for all models in our analysis. The rows represent different machine learning models, and the columns represent the features (i.e., variables) used in each model. The features were further stratified into clinical, histological, and genetic features. The colors in the heatmap reflect the importance scores, ranging from dark red (highest importance) to yellow (lowest importance). The importance scores were calculated using permutation feature importance, which is a technique that evaluates the importance of each feature by randomly permuting its values and measuring the impact on the model's performance. The resulting importance scores were then scaled between 0 and 1 for each model so that the scores are comparable across models. We can see that some features have consistently high importance across all models, while others have variable importance depending on the model. This suggests that some features may be more robust and informative for predicting survival outcomes than others, justifying the evaluation of the c-index for both all features and important features solely in Table 2.

## Discussion

The present study included multimodal data (genomics, pathology, and clinical features) for survival prediction in OSCC patients. Our results provide evidence of improved prediction capacity by incorporating more patient information in prediction tasks for survival prediction in OSCC patients.

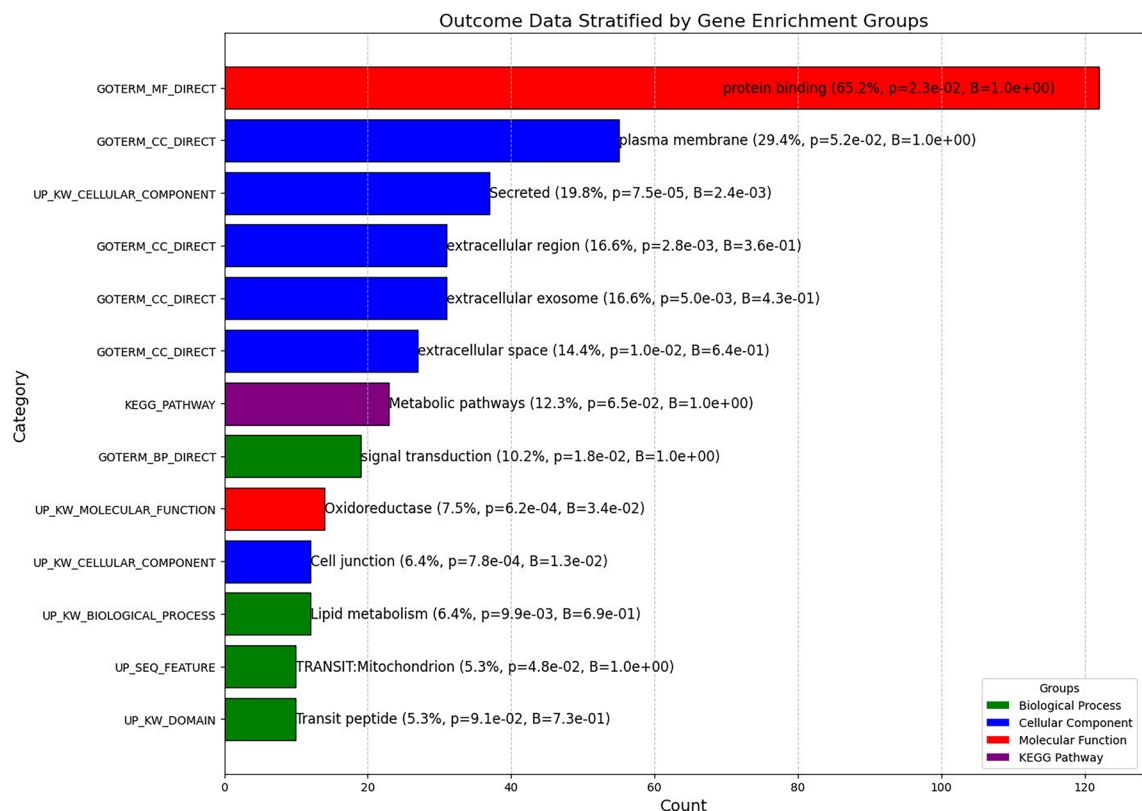
In this study, we employed a combination of sophisticated models, including the Cox Proportional Hazards (CPH) model implemented by `CoxPHSurvivalAnalysis` from `sksurv.linear_model`, as well as advanced machine learning models such as `RandomSurvivalForest` and `GradientBoostingSurvivalAnalysis` from `sksurv.ensemble`, `FastSurvivalSVM` from `sksurv.svm`, and `KerasRegressor` from `keras.wrappers.scikit_learn`. This approach aimed to leverage the strengths of traditional hazards-based models while also exploring the potential benefits of using more advanced machine learning and deep learning techniques for outcome prediction in cancer patients. While the traditional CPH model is useful for inferring the impact of variables on survival curves, integrating machine learning and deep learning methods can further enhance predictive accuracy. Artificial intelligence-driven approaches emphasize prediction over explanation and can address challenges like nonlinear



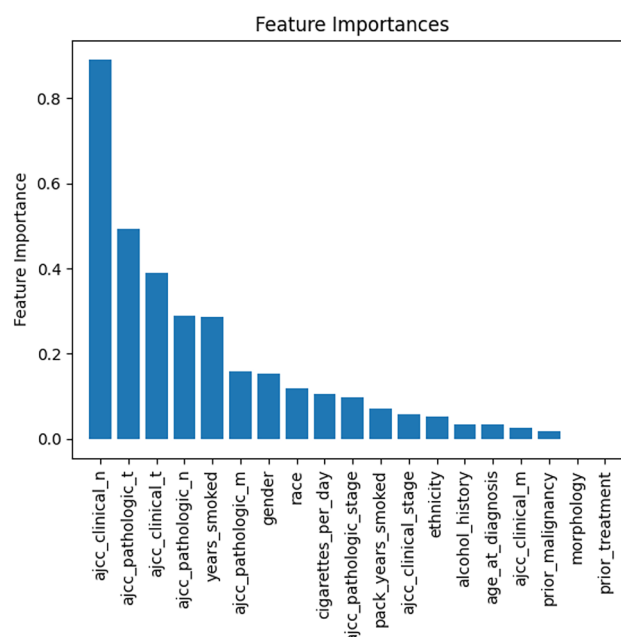
**Figure 4.** Heatmap of top differentially expressed genes ( $n=72$ ) as identified by ElasticNet. The columns represent the normal (red) and tumor solid tissue (black). The color scale ranges from blue (indicating low expression) to red (indicating high expression). The dendrogram for the rows (genes) of the heatmap represents the hierarchical clustering of genes based on their similarity in expression across the samples. The dendrogram for the columns (samples) of the heatmap represents the hierarchical clustering of samples based on their similarity in expression across the genes. The dendrograms show how similar or dissimilar the samples or genes are to each other based on their expression patterns. The height of the dendrogram represents the distance between clusters, with shorter distances indicating greater similarity or correlation. Clusters that are more similar are grouped together and have a common color in the heatmap. A total of 200 DEGs were analyzed further and are not shown here for better visualization. The figure was generated using Python (version 3.10.4).

gene interactions and multicollinearity, which may pose difficulties for conventional statistical methods. By examining extensive data, encompassing factors such as disease status, pathology, and genetic profiles, machine learning and deep learning models can determine the most advantageous treatment or clinical trial for a patient. Traditional statistical analyses may struggle with multicollinearity, particularly when integrating new prognostic factors. However, specific machine learning algorithms remain unaffected by significant collinearity among variables and can manage high-dimensional data<sup>30</sup>. For instance, Random Survival Forest (RSF) has outperformed classic CPH regressions in multiple studies<sup>31–33</sup>. Additionally, deep learning neural networks have demonstrated enhanced predictive accuracy compared to the traditional CPH model<sup>34–36</sup>. In a prior study, a nomogram predicting survival based on clinical variables and molecular markers for 68 oral SCC patients (validation dataset) achieved a c-index of 0.697, similar to the CoxPHSurvivalAnalysis result in this study<sup>37</sup>. Notably, RSF and deep learning models showed further improvements. The c-index serves as an excellent survival performance metric, as it is independent of a single fixed evaluation interval and considers censoring. The C-index's ability to handle censored data effectively is particularly pivotal in analyzing OSCC datasets, where such data is prevalent. Furthermore, its integration with our feature importance analysis, especially through the C-index reduction technique, enriches the interpretability and clinical applicability of our model. This approach, favoring the C-index over time-dependent AUC, aligns our work more closely with the practical demands and standards of clinical prognosis in OSCC. Our methodology showcases the potential to boost predictive accuracy in cancer patient outcomes beyond the capabilities of traditional statistical methods by employing a mix of advanced techniques.

Notably, there are several other techniques for multimodal data processing, and the present work applied only one of them (early fusion). In the field of multimodal fusion, prior research has investigated early and late fusion



**Figure 5.** Results of the gene enrichment analyses. Genes with a total count of  $\geq 10$  are shown. Gene Ontology (GO) terms and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways were used as functional annotation categories. The enrichment analysis was performed with the default parameters, and the significance threshold was set at  $p < 0.05$ . The resulting output included enriched terms, count (%), their corresponding  $p$ -values, and Benjamini-corrected  $p$ -values. The figure was generated using Python (version 3.10.4) based on data from KEGG<sup>23–25</sup>.



**Figure 6.** Feature importance analysis for the clinical variables. The figure was generated using Python (version 3.10.4).

		Clinical	Pathology	Genetics	Multimodal
All features	RSF	0.714	0.530	0.529	0.722
	GBSA	0.691	0.539	0.542	0.633
	FastSVM	0.684	0.489	0.527	0.625
	CoxPH	0.686	0.503	0.547	0.633
	DeepSurv	0.462	0.538	0.501	0.515
Important features	RSF	0.698	0.635	0.637	0.834
	GBSA	0.672	0.568	0.593	0.747
	FastSVM	0.706	0.500	0.636	0.718
	CoxPH	0.708	0.510	0.632	0.742
	DeepSurv	0.413	0.557	0.503	0.635

**Table 2.** Unimodal and multimodal artificial intelligence-based analyses for survival prediction. The values represent the c-index. The c-index is a commonly used metric in survival analysis that evaluates the predictive accuracy of a model. It measures the probability that, given two randomly selected patients, the patient with the worse prognosis, according to the model, will experience an event (such as death) before the patient with the better prognosis. A c-index of 0.5 indicates that the model is no better than a random chance at predicting outcomes, while a c-index of 1.0 indicates perfect predictive accuracy.



**Figure 7.** The heatmap displays the pooled feature importance scores for all models in our analysis. The rows represent different machine learning and deep learning models, and the columns represent the features (i.e., variables) used in each model. The features were further stratified into clinical, histological, and genetic features (separated by black vertical lines). The colors in the heatmap reflect the importance scores, ranging from dark red (highest importance) to yellow (lowest importance). The figure was generated using Python (version 3.10.4).

techniques. Early fusion concatenates features, while late fusion combines modalities through weighted averaging, failing to account for cross-modal interactions<sup>38,39</sup>. However, recent studies have demonstrated successful multimodal fusion through bilinear and graph-based models that exploit relationships within each modality<sup>40,41</sup>. Adversarial representation graph fusion (ARGF) has introduced a hierarchical interaction learning procedure, generating bimodal and trimodal interactions based on unimodal and bimodal dynamics<sup>42</sup>. Promising attempts have combined pathology and genomic data for cancer prognosis<sup>43,44</sup>. The Kronecker product, which creates a

high-dimensional feature of quadratic expansion based on pairings of two input feature vectors, has demonstrated superior cancer survival prediction<sup>40,45,46</sup>. However, it may introduce a large number of parameters, increasing computational costs and risking overfitting<sup>47,48</sup>. Hierarchical factorized bilinear fusion for cancer survival prediction (HFBSurv) integrates genomic and image features, overcoming these limitations<sup>49</sup>. Recently, PONET was proposed at a scientific conference. PONET is an innovative biological pathway-driven pathology-genomic deep learning model that combines pathological images and genomic information to enhance survival prediction and pinpoint genes and pathways responsible for varying survival rates among patients<sup>8</sup>. Future validation of this model will provide information about its usefulness in clinics.

Despite the promising results obtained in this study, there are some limitations that need to be addressed. First, our study is based on retrospective data from the TCGA dataset, which may limit the generalizability of our findings to other cohorts or populations. In addition, the sample size of our study is relatively small, which may limit the statistical power and generalizability of our results. Further studies with larger sample sizes are needed to validate our findings. Moreover, the multimodal data processing approach used in our study requires sophisticated algorithms and computational resources, which may limit its feasibility for routine clinical practice. However, with the rapid advancements in computing power and AI technologies, the feasibility and practicality of this approach may improve in the future. Finally, our study is limited to the use of genomic, pathology, and clinical data, and other data modalities, such as radiomics and proteomics, were not included in the analysis. Future studies that incorporate multiple data modalities may provide a more comprehensive understanding of the disease and improve the accuracy of prognostic prediction.

## Conclusions

In this study, we present an approach for predicting the survival of OSCC cancer patients using multimodal data processing techniques. We have applied a stratification method to distinguish unimodal and multimodal data processing with regard to evaluation metrics. By using a multimodal data fusion technique, we evaluated several model architectures across multiple data modalities. Our results demonstrate that the use of multimodal data processing techniques can significantly improve the accuracy of predictive algorithms, leading to more accurate long-term survival predictions for patients with OSCC. These hybrid algorithms are capable of leveraging the rich and complex information provided by multiple high-dimensional data modalities in precision medicine-based clinical practices. By providing clinicians with accurate and reproducible predictions of patient prognosis, these algorithms hold great promise for enhancing the management of cancer patients.

## Data availability

The codes and algorithm structures are available from: <https://github.com/Freiburg-AI-Research>. The raw data is publicly available from <https://portal.gdc.cancer.gov/> (TCGA-HNSC dataset). Gene enrichment analyses were conducted with data from Kyoto Encyclopedia of Genes and Genomes (KEGG)<sup>23–25</sup>.

Received: 10 July 2023; Accepted: 3 March 2024

Published online: 07 March 2024

## References

- Sung, H. *et al.* Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 Countries. *CA Cancer J. Clin.* **71**, 209–249 (2021).
- Chen, S.-H., Hsiao, S.-Y., Chang, K.-Y. & Chang, J.-Y. New insights into oral squamous cell carcinoma: From clinical aspects to molecular tumorigenesis. *Int J. Mol. Sci.* **22**, 2252 (2021).
- Adrien, J., Bertolus, C., Gambotti, L., Mallet, A. & Baujat, B. Why are head and neck squamous cell carcinoma diagnosed so late? Influence of health care disparities and socio-economic factors. *Oral Oncol.* **50**, 90–97 (2014).
- González-Moles, M. Á., Aguilar-Ruiz, M. & Ramos-García, P. Challenges in the early diagnosis of oral cancer, evidence gaps and strategies for improvement: A scoping review of systematic reviews. *Cancers* **14**, 4967 (2022).
- Russo, D. *et al.* Development and validation of prognostic models for oral squamous cell carcinoma: A systematic review and appraisal of the literature. *Cancers* **13**, 5755 (2021).
- Carreras-Torras, C. & Gay-Escoda, C. Techniques for early diagnosis of oral squamous cell carcinoma: Systematic review. *Med. Oral. Patol. Oral. Cir. Bucal.* **20**, e305–315 (2015).
- Alabi, R. O. *et al.* Machine learning in oral squamous cell carcinoma: Current status, clinical concerns and prospects for future-A systematic review. *Artif. Intell. Med.* **115**, 102060 (2021).
- Qiu L, Khormali A, & Liu K. Deep Biological Pathway Informed Pathology-Genomic Multimodal Survival Prediction. (2023) [cited 2023 Apr 3]; <https://arxiv.org/abs/2301.02383>
- Vale-Silva, L. A. & Rohr, K. Long-term cancer survival prediction using multimodal deep learning. *Sci. Rep.* **11**, 13505 (2021).
- Carrillo-Perez, F. *et al.* Machine-learning-based late fusion on multi-omics and multi-scale data for non-small-cell lung cancer diagnosis. *JPM* **12**, 601 (2022).
- Lipkova, J. *et al.* Artificial intelligence for multimodal data integration in oncology. *Cancer Cell.* **40**, 1095–1110 (2022).
- Steyaert, S. *et al.* Multimodal deep learning to predict prognosis in adult and pediatric brain tumors. *Commun. Med.* **3**, 44 (2023).
- Saravi, B. *et al.* Artificial intelligence-driven prediction modeling and decision making in spine surgery using hybrid machine learning models. *J. Personal. Med.* **12**, 509 (2022).
- Zuley, M.L., Jarosz, R., Kirk, S., Lee, Y., Colen, R., & Garcia, K., *et al.* The Cancer Genome Atlas Head-Neck Squamous Cell Carcinoma Collection (TCGA-HNSC), The Cancer Imaging Archive, 2016 (Accessed 3 Apr 2023); <https://wiki.cancerimagingarchive.net/x/VYG0>
- Li, X. *et al.* Multi-omics analysis reveals prognostic and therapeutic value of cuproptosis-related lncRNAs in oral squamous cell carcinoma. *Front. Genet.* **13**, 984911 (2022).
- Zou, C. *et al.* Identification of immune-related risk signatures for the prognostic prediction in oral squamous cell carcinoma. *J. Immunol. Res.* **2021**, 6203759 (2021).
- Macenko, M., Niethammer, M., Marron, J.S., Borland, D., Woosley, J.T., Xiaojun, G., *et al.* A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, 2009 (IEEE, accessed 4 Apr 2023). P. 1107–1110. <http://ieeexplore.ieee.org/document/5193250/>

18. Vahadane, A. *et al.* Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Trans. Med. Imaging* **35**, 1962–1971 (2016).
19. Salvi, M., Acharya, U. R., Molinari, F. & Meiburger, K. M. The impact of pre- and post-image processing techniques on deep learning frameworks: A comprehensive review for digital pathology image analysis. *Comput. Biol. Med.* **128**, 104129 (2021).
20. Carpenter, A. E. *et al.* Cell Profiler: Image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**, R100 (2006).
21. Hughey, J. J. & Butte, A. J. Robust meta-analysis of gene expression using the elastic net. *Nucleic Acids Res.* **43**, e79 (2015).
22. Tschodu, D. *et al.* Re-evaluation of publicly available gene-expression databases using machine-learning yields a maximum prognostic power in breast cancer. *Sci. Rep.* **13**, 16402 (2023).
23. Kanehisa, M. & Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
24. Kanehisa, M. Toward understanding the origin and evolution of cellular organisms. *Protein Sci.* **28**, 1947–1951 (2019).
25. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M. & Ishiguro-Watanabe, M. KEGG for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Res.* **51**, D587–D592 (2023).
26. Ye, H. *et al.* Metabolism-related bioinformatics analysis reveals that HPRT1 facilitates the progression of oral squamous cell carcinoma in vitro. *J. Oncol.* **2022**, 1–16 (2022).
27. Ferreira, A.-K. *et al.* Survival and prognostic factors in patients with oral squamous cell carcinoma. *Med. Oral. Patol. Oral. Cir. Bucal.* **26**, e387–e392 (2021).
28. Asio, J., Kamulegeya, A. & Banura, C. Survival and associated factors among patients with oral squamous cell carcinoma (OSCC) in Mulago hospital, Kampala, Uganda. *Cancers Head Neck.* **3**, 9 (2018).
29. Girod, A., Mosseri, V., Jouffroy, T., Point, D. & Rodriguez, J. Women and squamous cell carcinomas of the oral cavity and oropharynx: Is there something new?. *J. Oral Maxillof. Surg.* **67**, 1914–1920 (2009).
30. Wong, K., Rostomily, R. & Wong, S. Prognostic gene discovery in glioblastoma patients using deep learning. *Cancers* **11**, 53 (2019).
31. Hsich, E., Gorodeski, E. Z., Blackstone, E. H., Ishwaran, H. & Lauer, M. S. Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. *Circ. Cardiovasc. Qual. Outcomes* **4**, 39–45 (2011).
32. Ishwaran, H., Kogalur, U. B., Gorodeski, E. Z., Minn, A. J. & Lauer, M. S. High-dimensional variable selection for survival data. *J. Am. Stat. Assoc.* **105**, 205–17 (2010).
33. Ishwaran, H., Kogalur, U. B., Chen, X. & Minn, A. J. Random survival forests for high-dimensional data. *Stat. Anal. Data Min. ASA Data Sci. J.* **2011**(4), 115–32 (2011).
34. Katzman, J. L. *et al.* DeepSurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **18**, 187–202 (2018).
35. Sargent, D. J. Comparison of artificial neural networks with other statistical approaches. *Cancer* **91**, 1636–1642 (2001).
36. Xiang, A., Lapuerta, P., Ryutov, A., Buckley, J. & Azen, S. Comparison of the performance of neural network methods and Cox regression for censored survival data. *Comput. Stat. Data Anal.* **34**, 243–57 (2000).
37. Nie, Z., Zhao, P., Shang, Y. & Sun, B. Nomograms to predict the prognosis in locally advanced oral squamous cell carcinoma after curative resection. *BMC Cancer* **21**, 372 (2021).
38. Nojavanasghari, B., Gopinath, D., Koushik, J., Baltrušaitis, T., & Morency, L. P. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. 284–288 (2016).
39. Kampman, O., Barezi, E. J., Bertero, D., & Fung, P. Investigating audio, video, and text fusion methods for end-to-end automatic personality prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* vol. 2. 606–611 (2018).
40. Wang, Z., Li, R., Wang, M. & Li, A. Gpbn: Deep bilinear network integrating both genomic data and pathological images for breast cancer prognosis prediction. *Bioinformatics* **27**, 2963–2970 (2021).
41. Subramanian, V., Syeda-Mahmood, T., & Do, M. N. Multimodal fusion using sparse cca for breast cancer survival prediction. In *Proceedings of IEEE 18th International Symposium on Biomedical Imaging (ISBI)*. 1429–1432 (2021).
42. Mai, S., Hu, H., & Xing, S. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion. In *Proceedings of the AAAI Conference on Artificial Intelligence* 164–172 (2020).
43. Mobadersany, P. *et al.* Predicting cancer outcomes from histology and genomics. Predicting cancer outcomes from histology and genomics using convolutional networks. *Proc. Natl. Acad. Sci.* **115**, 2970–2979 (2018).
44. Wang, C. *et al.* A cancer survival prediction method based on graph convolutional network. *IEEE Trans. Nanobiosci.* **19**, 117–126 (2020).
45. Zadeh, A., Chen, M., Poria, S., Cambria, E., & Morency, L. P. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1103–1114 (2017).
46. Chen, R. J. *et al.* Pathomic fusion: An integrated framework for fusing histopathology and genomic features for cancer diagnosis and prognosis. *IEEE Trans. Med. Imaging* **41**, 757–770 (2022).
47. Kim, J. H., On, K. W., Lim, W., Kim, J., Ha, J. W., & Zhang, B. T. Hadamard product for low-rank bilinear pooling. In *Proceedings of International Conference on Learning Representations*, 1–14 (2017).
48. Liu, Z., Shen, Y., Lakshminarasimhan, V. B., Liang, P. P., Zadeh, A., & Morency, L. P. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. 2247–2256 (2021).
49. Li, R., Wu, X., Li, A. & Wang, M. Hfbsurv: Hierarchical multimodal fusion with factorized bilinear models for cancer survival prediction. *Bioinformatics* **38**, 2587–2594 (2022).

## Author contributions

Conceptualization: A.V., B.S., M.V., and V.S.; Data curation: A.V., M.V., and B.S.; Formal analysis of results and datasets: A.V., V.S., R.B., A.K., J.W., F.H., and G.L.; Investigation of further analyses to be conducted: V.S., R.B., S.H., A.K., J.W., F.H., S.C., and G.L.; Methodological conception: A.V., S.H., and B.S.; Resources for studies: A.K., J.W., F.H., S.C., and G.L. Validation of results: A.V., M.V., V.S., F.H., S.C., and G.L. Visualization of results: A.V., M.V., V.S., and B.S.; Writing—original draft, A.V., M.V., V.S., and B.S.; Writing—review and editing: S.H., R.B., A.K., J.W., F.H., S.C., and G.L. All authors revised and improved the manuscript and take accountability for the integrity and accuracy of the work.

## Funding

Open Access funding enabled and organized by Projekt DEAL. This research received no external funding.

## Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-024-56172-5>.

**Correspondence** and requests for materials should be addressed to A.V.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024