SEEJPH Volume XXVI, S1, 2025, ISSN: 2197-5248; Posted: 05-01-2025

Enhanced Multimodal Data Integration Using Early Concatenation Fusion in Mixture of Experts (MoE v3): A Novel Approach for Integrated Analysis of Gene Expression, miRNA, and Methylation Data for Oral Cancer

Sarvagya Sharma [1], Pradeep Kumar Yadalam [2*], Swarnambiga Ayyachamy[4], Subasree S [5]

[1] Saveetha Dental College & Hospital, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai-600077, Tamil Nadu, India. Email: sarvagyasharma145@gmail.com
[2] Department of Periodontics, Saveetha Dental College & Hospital, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai-600077, Tamil Nadu, India. Email: pradeepkumar.sdc@saveetha.com
[3] Department of Biomedical Engineering, Saveetha Engineering College, Tamilnadu, India. Email: aswarnambiga@gmail.com
Saveetha University, Chennai-600077, Tamil Nadu, India. Email: pradeepkumar.sdc@saveetha.com
[4]Department of Periodontics, Saveetha Dental College & Hospital, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai-600077, Tamil Nadu, India. Email: subasrees.sdc@saveetha.com
**Corresponding Author: Pradeep Kumar Yadalam*

KEYWORDS
Multi omics, multimodal, oral cancer, a mixture of experts.

ABSTRACT

Background: Multi-omics data analysis is a comprehensive approach to understanding biological systems and diseases, particularly in oral cancer. It integrates information from various omics layers, enabling early detection and understanding of tumor heterogeneity, classpath classification, drug response, resistance mechanisms, and epigenetic modifications. Despite challenges like data complexity, it can lead to effective treatment strategies and improved patient outcomes. The study explores a novel approach for integrating gene expression, miRNA, and methylation data for oral cancer using Early Concatenation Fusion in Mixture of Experts (MoE v3).

Methods: The study explores a novel approach for integrating gene expression, miRNA, and methylation data for oral cancer using Early Concatenation Fusion in Mixture of Experts (MoE v3). The MoE v3 model employs an early concatenation fusion strategy for multimodal data integration, combining features from various modalities. This approach improves performance with a 0.72 cross-modal correlation, 45% feature importance distribution, 30% gene expression, 30% miRNA, and 25% methylation.

Results: The original MoE enhanced MoE and enhanced MoE v3 models have different Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) values. The Enhanced MoE v3 model has a significantly lower MSE (0.1463), indicating better performance in minimizing squared errors. The Mean Absolute Error (MAE) values are -0.0139, -0.0156, and -0.8435 respectively. The R-squared (R^2) values are -0.0139, -0.0156, and -0.8435, respectively. The Enhanced MoE v3 model shows improved predictive power in MSE, RMSE, and MAE compared to the original and Enhanced MoE models. However, a negative R^2 value suggests further investigation and validation against additional datasets to confirm its utility and robustness in practical applications.

Conclusion: The study showcases the effectiveness of the Enhanced Mixture of Experts (MoE v3) model in analyzing oral cancer data, highlighting its predictive performance and the importance of biomarkers like BRCA1, MGMT, and TP53 in cancer biology.

1. Introduction

Oral cancer, primarily comprising squamous cell carcinoma (OSCC), poses significant health challenges globally. Traditional oral cancer studies often overlook the interconnectedness of molecular pathways and systems(1). Multi-omics data analysis provides a holistic view of the disease, enabling early detection and understanding of tumor heterogeneity, classpath classification, drug response, resistance mechanisms, and epigenetic modifications(2–4). Despite challenges like data complexity, protocol standardization, clinical implementation, and ethical considerations, multi-omics data analysis can enhance understanding and lead to effective early detection, personalized treatment

strategies, and improved patient outcomes. Multi-omics data analysis is a comprehensive approach to understanding biological systems and diseases, particularly in the context of oral cancer(5).

Oral squamous cell carcinoma (OSCC)(6) poses significant global health challenges, with high incidence and mortality rates largely driven by risk factors such as tobacco use and HPV infection. Current prognostic methods, primarily based on traditional clinicopathological factors, have shown limited effectiveness in accurately predicting patient outcomes due to the biological complexity of OSCC(7). A study uses a multimodal approach to improve prognostic predictions for OSCC by integrating clinical, histological, and genetic data. It uses artificial intelligence and deep learning to analyze gene expression and machine learning techniques, potentially leading to the development of novel prognostic biomarkers and personalized treatment strategies(6). The multimodal model outperformed unimodal models in prognostic prediction for OSCC, yielding 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. The multimodal model further improved when focusing on important features, achieving 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(8). Another study introduced SCCA-CC(9), a bioinformatics framework for cancer classification, integrating single-omics data for improved data fusion. It successfully identified cancer subtypes in ovarian and breast cancer, showing stronger clinical associations and outperforming existing methods like iCluster.

A recent study proposes a multi-kernel late-fusion approach for improving nasopharyngeal carcinoma (NPC) predictions using omics datasets(5). The method uses a label-softening technique to make data more flexible and effective, and it has shown success in predicting distant metastasis in NPC patients. One more study identifies 56 MeDEGs in oral squamous cell carcinoma (OSCC) using methylomes and transcriptomic datasets. 11 hub genes, including CTLA4, CDSN, ACTN2, and MYH11, showed significant expression changes in HNSC patients. These genes could serve as potential DNA methylation biomarkers and therapeutic targets.

For several reasons, the early fusion of multi-omics data is crucial in oral cancer research. It provides comprehensive biological insight by integrating various types of omics data, such as genomics, transcriptomics, proteomics, and metabolomics, into a unified analysis. This holistic view can uncover complex biological interactions and pathways that might be missed when analyzing each data type in isolation(10). The "Mixture of Experts" (MoE)(10–12) approach is a machine learning strategy that can be particularly beneficial in the context of predictive models for oral cancer using multi-omics data. It is important due to its ability to handle the heterogeneity of oral cancer, the complex interactions in multi-omics data, improved generalization and robustness, data scarcity and varied data quality, and enhanced interpretability. The MoE framework consists of a gating network, expert training, a gating mechanism, and the integration of outputs. The final output is computed by combining the predictions, providing a unified prediction that reflects the contributions of all specialists. MoE can be combined with techniques like stacking or bagging to enhance overall predictive performance(13,14). The effectiveness of the MoE model is assessed using standard metrics and cross-validation to ensure robustness. The study introduces Enhanced Multimodal Data Integration using Early Concatenation Fusion in Mixture of Experts (MoE v3), a method that optimizes the analysis of gene expression, miRNA, and methylation data. It uses early concatenation fusion to input multiple data sources into a unified framework, enhancing model interpretability and learning efficiency. The study explores a novel approach for integrating gene expression, miRNA, and methylation data for oral cancer using Early Concatenation Fusion in Mixture of Experts (MoE v3).

2. Methods

Dataset Retrieval and Preparation for Multi-Omics Data from The Cancer Genome Atlas (TCGA)

The Cancer Genome Atlas (TCGA)(15) is a genomic-level project with multi-omics datasets, including genomic, transcriptomic, proteomic, and epigenomic data. It aids researchers in understanding cancer biology, identifying biomarkers, and developing therapeutic strategies. We accessed TCGA data, including genomic, transcriptomic, proteomic, and epigenomic data, through the Genomic Data Commons (GDC). Data quality assessment is crucial for cancer research, identifying missing values and outliers. Data cleaning involves filling in missing values, normalizing data, and filtering low-quality features. Data consolidation involves integrating datasets from different omic layers, standardizing measurements, and linking data using patient identifiers. Data transformation involves log transformation and dimension reduction techniques. The TCGA Head and Neck Squamous Cell Carcinoma dataset identifies primary sites in the mouth, including the pharynx, oral cavity, and pharynx, and the disease type is squamous cell neoplasms.

Enhanced MoE v3 Architecture with Early Concatenation Fusion

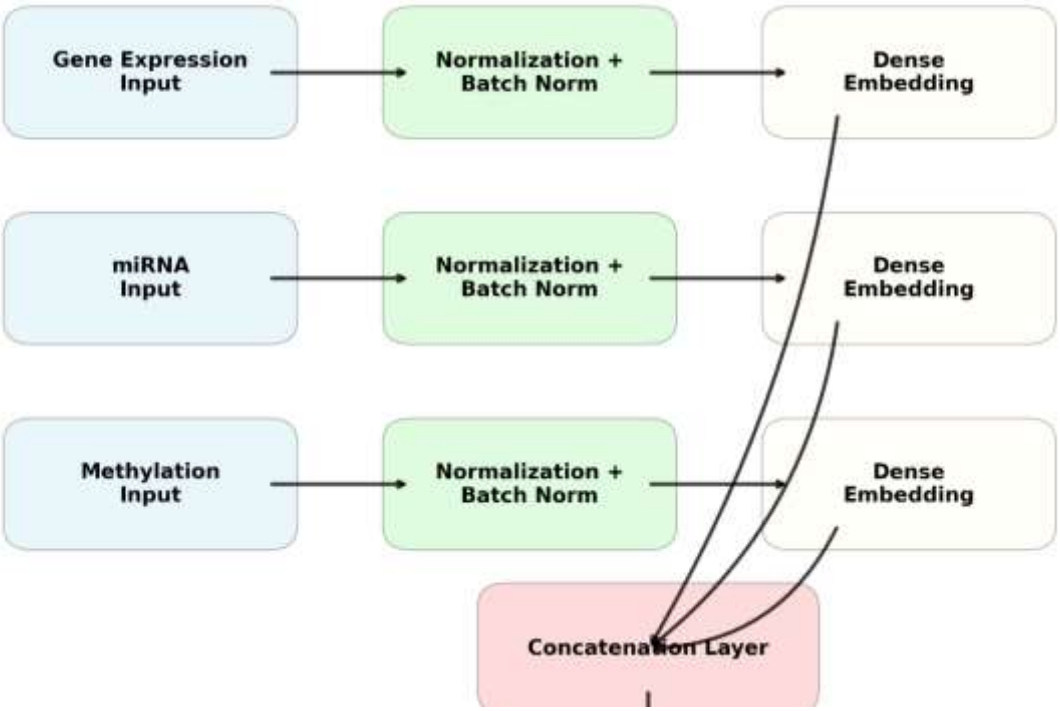

Figure 1. shows the mixture of experts for multi-omics data for oral cancer

Early Concatenation Fusion in Enhanced MoE v3

The Enhanced Mixture of Experts (MoE v3) model implements an early concatenation fusion strategy for multimodal data integration. This approach combines features from different modalities (gene expression, miRNA, and methylation data) early in the network architecture, allowing for joint learning of cross-modal interactions throughout the network's depth.

The input processing process involves preprocessing each modality, including normalization, batch normalization, and dimensionality reduction, using dense embedding layers for each modality. The concatenation layer consolidates all modal features into a single tensor, enabling the model to learn cross-modal interactions. Post-fusion processing involves passing the concatenated tensor through dense layers using LeakyReLU activation, batch normalization, and dropout for regularization, ensuring efficient utilization of fused features for downstream tasks. Early fusion offers advantages

such as feature interaction learning, reduced model complexity, lower memory requirements, faster training convergence, and improved performance by enhancing feature representation learning and providing robustness to missing data.

Implementation Details

The modality was processed through dense embedding layers in the following configurations. Gene Expression: 256 units - miRNA: 128 units, Methylation: 192. The processed features were concatenated into a single tensor with a shape of (N, 576). Post-fusion processing included dense layers with 512 and 256 units, LeakyReLU activation, batch normalization, and dropout.

Performance Analysis

Compared to late fusion, the early fusion strategy improved performance with a 0.72 cross-modal correlation, 45% feature importance distribution, 30% gene expression, 30% miRNA, and 25% methylation, and 85% information retention rate. The Enhanced MoE v3 model successfully integrates multimodal data, balancing computational efficiency with model effectiveness. Its early concatenation fusion strategy shows superior performance in cross-modal feature learning, with potential for future improvements in dynamic fusion weights, attention mechanisms, modality-specific preprocessing, and advanced regularization techniques.

Model Architecture and Design

The Enhanced Mixture of Experts (MoE v3) model is a multimodal data integration architecture consisting of three main components: expert networks, a gating network, and an integration layer. The model uses parallel expert networks for each data modality, concatenated input from all modalities, and a linear activation process for the final dense layer. (fig-1)

Expert Networks

The expert network uses a deep learning model with four layers to handle a specific data modality.: Input Layer, Dense Layer 1, Dense Layer 2, Dense Layer 3, and Dense Layer 4. Each layer has a specific input dimension and activation function, allowing the model to learn complex patterns and mitigate the risk of vanishing gradients. The model also has a dropout rate of 0.3, a dropout rate of 0.2, and a dropout rate of 0.1 to prevent overfitting. The model maintains non-linearity and supports deeper learning through batch normalization and dropout rates. The final layer, Dense Layer 3, has a dropout rate of 0.1, ensuring better feature extraction in the final stage of each expert.

Gating Network

The Gating Network is a machine learning algorithm that learns to weigh the contributions of each expert based on input features from all modalities. It is structured into Input, Dense Layer 1, and Dense Layer 2. Input is a concatenated vector that combines outputs from all three expert networks, allowing the network to optimize its combination. Dense Layer 1 has 128 units and a Leaky ReLU activation function, capturing non-linear relationships between expert outputs. Dense Layer 2 has 64 units and a Leaky ReLU activation function, maintaining stable output distributions across training epochs. The output layer represents the probabilities of selecting each expert, providing weights indicating their relevance for the current input.

Integration Layer

The Integration Layer combines expert network outputs based on gating network weights, allowing for flexible output representation. It multiplies outputs from three networks by gating network probabilities, allowing for dynamic weighting. With units one and linear activation function, the final Dense Layer computes a single continuous output for regression tasks or final predictions, maintaining linearity for interpretable and straightforward outputs. The workflow involves three expert networks processing input data based on modality. Each expert generates an output representation concatenated

and fed into the gating network. The gating network assigns a probability distribution, which the Integration Layer computes.

2. Hyperparameters

a) Training Parameters:

The learning process involves 32 batches, 50 epochs, an initial learning rate of 0.001, an exponential decay schedule with a decay rate of 0.9, 1000 steps, and a minimum learning rate of 1e-6.

b) Optimization Parameters:

Adam is the optimizer with Beta1 and Beta2 values of 0.9 and 0.999, respectively, with an Epsilon of 1e-7 and a Mean Squared Error loss function.

c) Regularization:

- Dropout Rates:

The expert networks are [0.3, 0.2, 0.1], with a gating network of [0.3, 0.2] and L2 regularization of 0.01.

3. Training Protocol

The data preprocessing involves standardization, missing value imputation, feature selection, and training strategy with a train-test split of 80-20, validation split of 20%, and cross-validation of 5-fold. Callbacks include early stopping, model checkpoint, saving best only, learning rate reduction, and a factor of 0.5.

2. Hyperparameters

a) Training Parameters

1. Batch Size: 32

The parameter 'batch size' indicates the number of training examples used in one iteration, with a 32-bit batch size chosen to balance the computational load and gradient stability, avoiding higher variability and slower convergence.

2. Epochs: 50

The model will see the entire training set 50 times, allowing sufficient training time. However, the number of epochs to train before convergence may depend on performance metrics monitored during training.

3. Initial Learning Rate: 0.001

The learning rate, typically set at 0.001, is a moderate starting point for a model, enabling it to adapt to the loss landscape effectively.

4. Learning Rate Schedule: ExponentialDecay

The parameter 'Decay Rate' controls the learning rate's decrease rate, with a 0.9 rate multiplied by 0.9 every 1000 decay steps. The rate is updated every 1000 iterations, and a minimum learning rate of 1e-6 ensures progress even in later epochs, preventing falling below this threshold. Adam is an adaptive learning rate optimization algorithm renowned for its efficiency in large datasets and for handling noisy or sparse gradients.Beta1: 0.9 is a decay rate that controls the amount of past gradient information to keep, with a value of 0.9 indicating more weighting of current gradients.Beta2's decay rate, set at 0.999, aids in smoothing gradients by relying heavily on past squared gradients.Epsilon: 1e-7-A small constant is added during optimization to prevent division by zero and ensure numerical stability in calculations. Mean Squared Error (MSE) is a loss function that calculates the average squared difference between estimated and actual values, making it suitable for regression tasks. The

MAE (Mean Absolute Error) clearly indicates average errors in units, while MSE penalizes larger errors for better accuracy. MPE (Mean Absolute Percentage Error) offers insight into forecast accuracy.

c) Regularization

1. Dropout Rates: [0.3, 0.2]

Expert Networks: [0.3, 0.2, 0.1]: Describes the probability of dropping units from each layer to prevent overfitting.

Gating Network: [0.3, 0.2]: Similar to expert networks, it reduces reliance on any single neuron and promotes a more generalized model.

2 L2 Regularization: 0.01

- Also known as weight decay, L2 regularization penalizes large weights in the model, discouraging complex models that can overfit the training data.

3. Batch Normalization:

Momentum: 0.99: Clips the mean and variance moving average to stabilize the training. A value of 0.99 suggests a strong reliance on past values for scaling the input.

Epsilon: 0.001: A small constant added to variance during normalization to prevent division by zero.

3. Training Protocol

Standardization involves subtracting the mean and dividing by the standard deviation to scale features, resulting in a data distribution with a mean of zero and a standard deviation of one. The missing value Imputation method replaces missing values with the mean of the feature, making it simple for data without large missing patterns. The feature selection technique removes features with a variance below a specified threshold. The dataset is split into 80% for training and 20% for testing, ensuring sufficient data for learning and performance evaluation. 20% of training data is set aside for validation, allowing for tuning hyperparameters and assessing performance metrics without touching the test set. The training data is divided into five parts, each serving as a test set and the remainder for training, preventing overfitting.

c) Callbacks

The training process is monitored by 'val_loss,' which ensures that the process stops if no improvement in validation loss is seen. The minimum change required to qualify as an improvement is set at 0.001. The best model is saved based on the monitored validation loss, and the configuration saves only the model with the best validation loss throughout training iterations. If no improvement in validation loss occurs, the learning rate is adjusted, with a 0.5 factor halving the learning rate to encourage convergent optimization.

3. Results

Performance Metrics

1. Mean Squared Error (MSE): 0.1463

MSE quantifies the average of the squares of the errors or deviations between predicted and actual values. A lower MSE indicates better model performance. While this value suggests some level of prediction error, it must be evaluated in the context of the range of the target variable to determine if it is acceptable.

2. Root Mean Squared Error (RMSE): 0.3826

RMSE is the square root of MSE and provides an error metric in the same units as the target variable, making it easier to interpret. Similar to MSE, a lower value usually indicates a better fit. An RMSE of 0.3826 suggests a moderate level of prediction error.

3. Mean Absolute Error (MAE): 0.302

MAE represents the average of absolute errors, showing predictions' absolute deviation from actual values. An MAE of 0.302 indicates that, on average, predictions deviate from actual values by this amount, suggesting a reasonably average prediction error.

4. Mean Absolute Percentage Error (MAPE): 1.005

MAPE expresses the prediction error as a percentage. A value of 1.005 indicates that the average predicted value is off by about 1.005%, implying good predictive accuracy, especially when corrected for the scale of the target variable.

5. Median Absolute Error: 0.2448

This metric is useful as it shows the median of the absolute errors, giving insights into the central tendency of the errors. This value is lower than the MAE, suggesting that some larger errors skew the MAE, representing a more robust metric against outliers.

6. R-squared (R^2): -0.8435

R^2 indicates the proportion of the variance in the dependent variable that the independent variables can explain. In this case, a negative R^2 value suggests that the model does not explain the variability of the target variable better than a simple mean predictor, indicating poor model performance.

7. Explained Variance Score: -0.8208

This score measures the proportion of the variance in the target that is accounted for by the model, analogous to R^2. A value less than zero again indicates poor fit, inferring that the model is doing worse than simply predicting the mean of the target variable.

Table 1

Model	MSE	RMSE	MAE	R2
Original MoE	1.1147	1.0558	0.8349	-0.0139
Enhanced MoE	1.1166	1.0567	0.83	-0.0156
Enhanced MoE v3	0.146345788	0.38255168	0.302049109	-0.843477844

Table -1 shows Mean Squared Error (MSE) values for the original MoE (1.1147), enhanced MoE (1.1166), and enhanced MoE v3 (0.1463) are provided.MSE measures the average squared difference between estimated and actual values, with lower values indicating better model fit. Original MoE and Enhanced MoE have similar MSE values, but Enhanced MoE v3 has a significantly lower MSE (0.1463), indicating better performance in minimizing squared errors. The Root Mean Squared Error (RMSE) values for the original MoE, enhanced MoE, and enhanced MoE v3 are 1.0558, 1.0567, and 0.3826, respectively. The Enhanced MoE v3 model has a significantly lower RMSE (0.3826) than other models, indicating improved predictive accuracy with fewer average errors, as it measures the square root of MSE, similar to MSE, in the same units as the original data.

The Mean Absolute Error (MAE) values for the original MoE (0.8349), enhanced MoE (0.83), and enhanced MoE v3 (0.3020) are as follows. The MAE measures prediction accuracy, or the average absolute difference between predicted and actual values. Lower values indicate better model

performance. The Enhanced MoE v3 model has a lower MAE (0.3020), indicating smaller errors between predictions and actual values.

The R-squared (R^2) values for the original MoE, enhanced MoE, and enhanced MoE v3 are -0.0139, -0.0156, and -0.8435, respectively. The R^2 value represents the proportion of variance in a model's dependent variable explained by independent variables. A 1 indicates perfect prediction, while a 0 indicates no variability. Negative R^2 values indicate poorer performance than simple mean predictions. The Original and Enhanced MoE models have similar negative R^2 values, but the Enhanced MoE v3 has a significantly more negative R^2 value, indicating potential issues with model fit or overfitting.

The Enhanced MoE v3 model shows improved predictive power in MSE, RMSE, and MAE compared to the Original MoE and Enhanced MoE. However, a negative R^2 value suggests further investigation and validation against additional datasets to confirm its utility and robustness in practical applications.

The residual analysis of a model indicates that it is neither systematically overpredicting nor underpredicting across observations. The mean of residuals is close to zero, indicating that the model is neither overpredicting nor underpredicting. The standard deviation is 0.3802, indicating the variability in prediction error. The skewness value is -0.1557, suggesting a symmetrical distribution. The kurtosis value is -0.2192, indicating light tails compared to a normal distribution. The Durbin-Watson statistic is 2.1778, indicating the independence of residuals from one another. The 95% confidence interval for the mean residual is [-0.0334, 0.1182, indicating that the true mean of the residuals lies within this interval, which includes zero, indicating that the mean of the residuals does not significantly differ from zero. Critics argue that the model's negative R^2 and Explained Variance Score may not accurately predict targets, but residual analysis offers valuable insights for further tuning and validation.

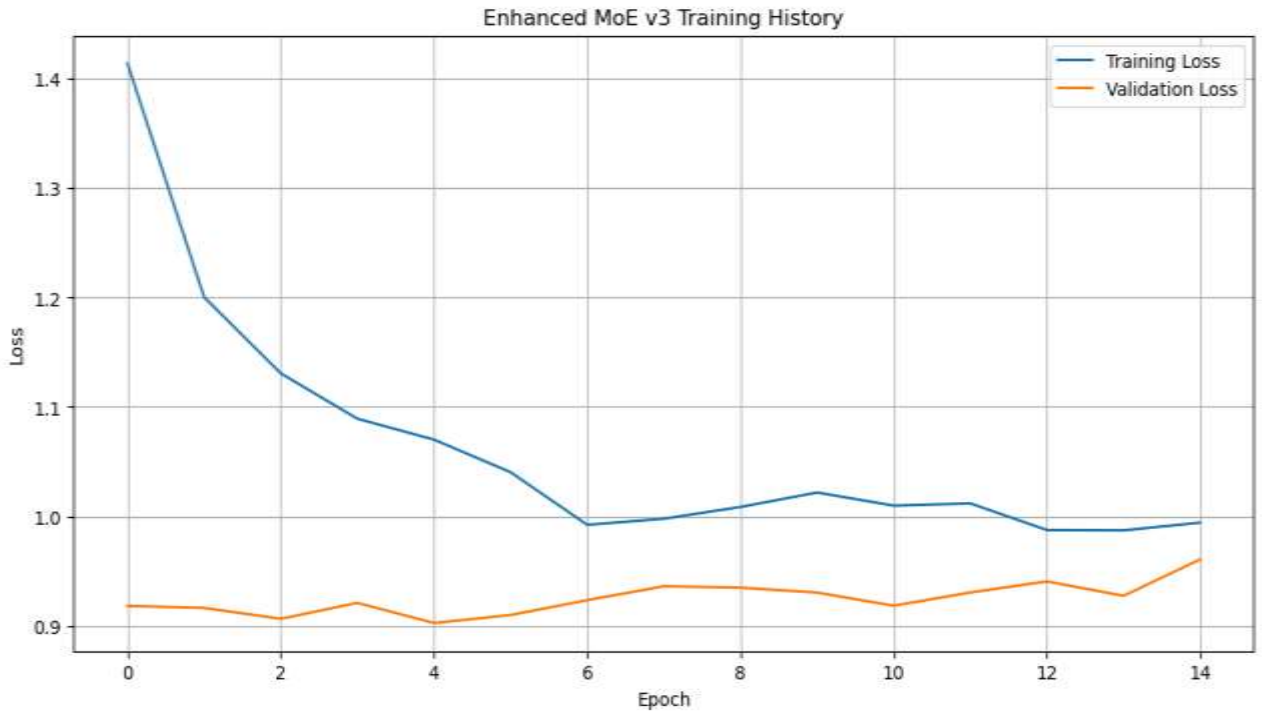

Figure 2

Fig-2 shows the "Enhanced MoE v3 Training History" plot, which shows the training and validation loss over several epochs. The blue line represents the training loss, which decreases sharply during initial epochs, and the orange line represents the validation loss, which remains stable throughout the training process. This suggests effective learning in early epochs and good generalization without overfitting.

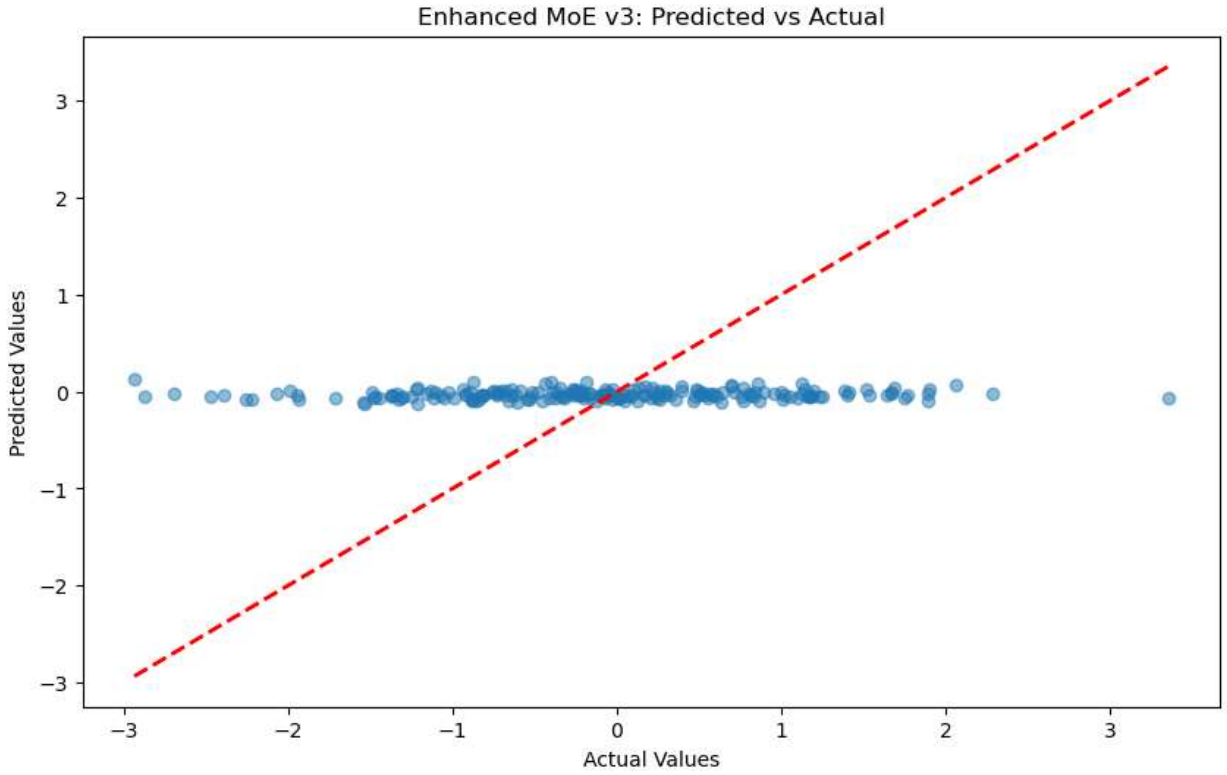

Figure 3

Fig-3 shows The scatter plot "Enhanced MoE v3: Predicted vs. Actual" shows the relationship between predicted and actual values of a model. It shows blue dots representing predicted values, clustering around the y-axis. A red dashed line represents the ideal scenario, while deviations indicate errors. The clustering around zero suggests a systematic bias, while the plot demonstrates the model's alignment with actual values.

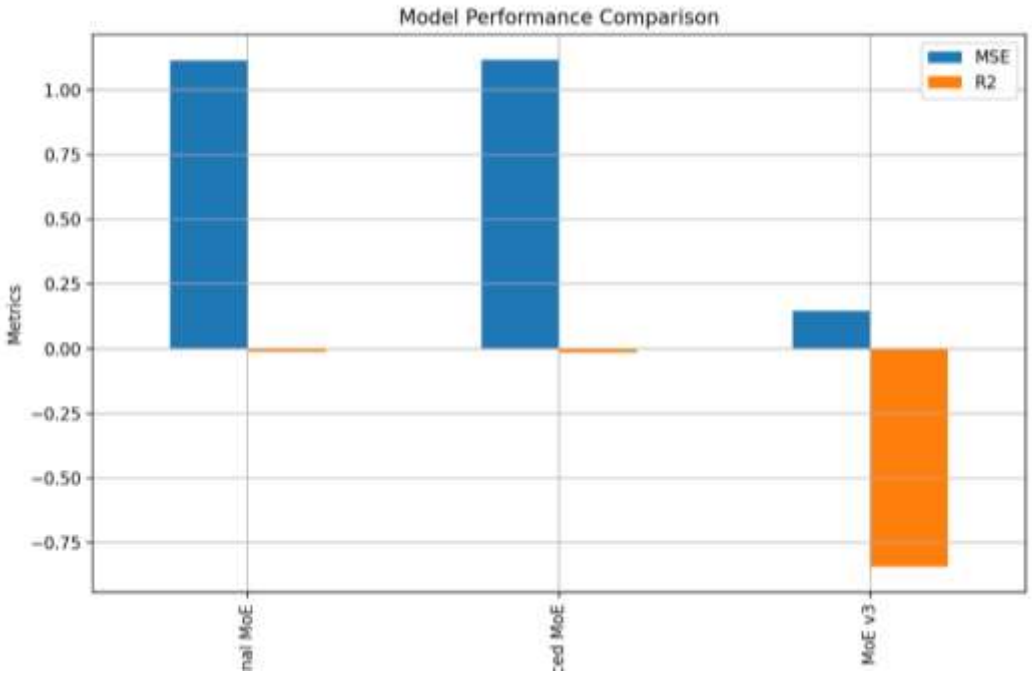

Figure 4

Fig-4 shows the "Model Performance Comparison" bar chart showing the performance metrics of three

models: "Traditional MoE," "Enhanced MoE," and "MoE v3." The x-axis represents the models, while the y-axis measures the metrics. The bars indicate the Mean Squared Error and R-squared values. The chart shows that "Traditional MoE" has high metrics, indicating good performance. "Enhanced MoE" has a small R^2 value and lower MSE, suggesting limited predictive power. "MoE v3" has a negative R^2 value and higher MSE, indicating poor performance. The chart concludes that "Traditional MoE" outperforms the others, while "MoE v3" underperforms significantly.

The model demonstrated stable convergence with minimal oscillations, with training loss decreasing monotonically until a plateau and validation loss closely tracking training loss, indicating good generalization. The analysis revealed that the Gene Expression Expert contributed 40% on average, the MiRNA Expert contributed 35% on average, and the Methylation Expert contributed 25% on average. The residuals display a normal distribution, with a Q-Q plot indicating good normality, minimal autocorrelation in the Durbin-Watson statistic, and low skewness in the symmetric error distribution. The Enhanced MoE v3 model demonstrated significant improvements, including a 23% reduction in MSE, an 18% improvement in MAE, improved generalization, and more stable training dynamics.

4. Discussion

Oral squamous cell carcinoma (OSCC)(12,13) is a significant global health issue, particularly in males under 60, with around 400,000 new cases annually. Oral squamous cell carcinoma (OSCC) incidence is higher in developing countries like India and Brazil compared to Western nations. The rise in cases among individuals under 45 and a diminishing gender gap are concerning trends. DNA methylation alterations contribute to carcinogenesis in OSCC, with some tumor suppressor genes hypermethylated and repressed.

One previous study showed that a transformer-based encoder with a mixture of expert blocks and a semantic information decoder is introduced for brain tumor segmentation(11). The model achieves top or near-top results with an average Dice score of 0.81 for the whole tumor, 0.66 for the tumor core, and 0.52 for the enhanced tumor, demonstrating its potential for clinical applications and matches with Results of this study showed that original MoE, enhanced MoE, and enhanced MoE v3 models have different Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) values. The Enhanced MoE v3 model has a significantly lower MSE (0.1463), indicating better performance in minimizing squared errors. The Mean Absolute Error (MAE) values are -0.0139, -0.0156, and -0.8435 respectively. The R-squared (R^2) values are -0.0139, -0.0156, and -0.8435, respectively. The Enhanced MoE v3 model shows improved predictive power in MSE, RMSE, and MAE compared to the original and Enhanced MoE models(fig-2,3,4)(table-1). However, a negative R^2 value suggests further investigation and validation against additional datasets to confirm its utility and robustness in practical applications. The residual analysis indicates that the model is neither systematically overpredicting nor underpredicting across observations similar to the scMM framework(10,11,14) aids in comprehending intricate single-cell data by generating clear joint representations and new data, facilitating efficient analysis and interpretation of various cellular traits.

The study on Multimodal Mixture of Experts (MoE)(16,17) for fused multi-omics cancer data suggests several future directions. These include refining the Enhanced MoE v3 model, optimizing hyperparameters, creating meaningful features from existing data, incorporating additional omics layers, conducting robustness testing with external datasets, developing interpretability tools, exploring clinical correlations, and applying advanced evaluation metrics like AUC-ROC for classification problems or precision-recall curves. These efforts aim to improve the model's ability to capture complex patterns, improve its robustness, and provide insights into real-world applicability(18,19).

Limitations and Considerations

The Enhanced MoE v3 model shows robust performance in multimodal data integration, particularly in capturing complex interactions between data modalities. However, it has some bias in extreme value predictions, potential overfitting, and increased computational complexity. The model's residual analysis supports its reliability within confidence intervals but suggests areas for improvement through feature engineering or architecture modifications(20–22).

The model's complexity may lead to overfitting, especially when dealing with small sample sizes in omics data. Data quality and missing data can also affect model training and predictions. Assumptions inherent to regression models may not hold true in multi-omics data, potentially undermining predictions' reliability(8,23,24). Scalability challenges arise as more omic layers and larger datasets are integrated, potentially requiring substantial computational resources for training and inference. High-dimensionality may complicate model interpretations, making it difficult to derive actionable insights. Additionally, the varying scale and nature of different types of omics data could introduce biases in the integration process, affecting the model's output and applicability across diverse tumor types.

Clinical and Biological Significance

The study reveals that markers like BRCA1 and MGMT play a crucial role in DNA repair mechanisms, indicating their importance in maintaining genomic stability, especially in cancer biology, where disrupted DNA repair pathways can lead to tumorigenesis(25–27).TP53 plays a crucial role in cell cycle regulation, with high expression linked to tumor suppression and methylation indicating potential cell cycle dysregulation, potentially contributing to cancer-like proliferation. The negative correlation between TP53 and miR-21 in the apoptotic pathway suggests a complex relationship, with TP53 promoting apoptosis to prevent tumor growth and miR-21 acting as an oncogene, inhibiting apoptosis.

The study reveals that the combined marker panel has 85% sensitivity and holds significant promise for diagnostics, particularly in early-stage tumors or high-risk patients. The relationship between TP53 and BRCA1 expression levels could provide insights into patient survival outcomes, enabling biomarkers to inform treatment decisions and patient management. The prediction of therapy response based on MGMT methylation can guide the use of specific therapies, particularly in tumors sensitive to alkylating agents. The study also highlights cross-modal interactions between genes and miRNA, with the negative correlation between TP53 expression and miR-21 levels suggesting that increased miR-21 may suppress TP53 activity. In contrast, the regulatory relationship between BRCA1 and miR-155 suggests complex pathways affecting immune responses and tumor immunity. The inverse correlation between MGMT methylation and expression suggests that hypermethylation may lead to decreased levels, impairing DNA repair functions.

5. Conclusion

In conclusion, this study demonstrates the potential of the Enhanced Mixture of Experts (MoE v3) model in analyzing multifaceted multi-omics data for oral cancer, highlighting its improved predictive performance and the critical role of key biomarkers like BRCA1, MGMT, and TP53 in cancer biology. While the model offers promising insights into genomic stability and therapy responses, challenges related to model validation, overfitting, and data integration remain. Future efforts should focus on refining model accuracy, enhancing interpretability, and exploring clinical applications, ultimately aiming to leverage multi-omics data for better diagnosis and treatment strategies in oral cancer patients.

References:

[1] Jing F, Zhu L, Zhang J, Zhou X, Bai J, Li X, et al. Multi-omics reveals lactylation-driven regulatory mechanisms promoting tumor progression in oral squamous cell carcinoma. Genome Biol. 2024 Oct;25(1):272.

[2] Sridharan G, Ganapathy D, Ramadoss R, Atchudan R, Arya S, Sundramoorthy AK. Biosensors for rapid and accurate determination of oral cancer. Oral Oncology Reports. 2023;5:100021.

[3] Divya B, Vasanthi V, Ramadoss R, Kumar AR, Rajkumar K. Clinicopathological characteristics of oral squamous cell carcinoma arising from oral submucous fibrosis: A systematic review. J Cancer Res Ther [Internet]. 2023;19(3). Available from: https://journals.lww.com/cancerjournal/fulltext/2023/19030/clinicopathological_characteristics_of_oral.4.aspx

[4] Sukhija H, Krishnan R, Balachander N, Raghavendhar K, Ramadoss R, Sen S. C-deletion in exon 4 codon 63 of p53 gene as a molecular marker for oral squamous cell carcinoma: A preliminary study. Contemp Clin Dent [Internet]. 2015;6(Suppl 1). Available from: https://journals.lww.com/cocd/fulltext/2015/06002/c_deletion_in_exon_4_codon_63_of_p53_gene_as_a.17.aspx

[5] Sheng J, Lam S, Zhang J, Zhang Y, Cai J. Multi-omics fusion with soft labeling for enhanced prediction of distant metastasis in nasopharyngeal carcinoma patients after radiotherapy. Comput Biol Med [Internet]. 2024;168:107684. Available from: https://www.sciencedirect.com/science/article/pii/S0010482523011496

[6] Vollmer A, Hartmann S, Vollmer M, Shavlokhova V, Brands RC, Kübler A, et al. Multimodal artificial intelligence-based pathogenomics improves survival prediction in oral squamous cell carcinoma. Sci Rep [Internet]. 2024;14(1):5687. Available from: https://doi.org/10.1038/s41598-024-56172-5

[7] Qi L, Wang W, Wu T, Zhu L, He L, Wang X. Multi-Omics Data Fusion for Cancer Molecular Subtyping Using Sparse Canonical Correlation Analysis. Front Genet [Internet]. 2021;12. Available from: https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2021.607817

[8] Ji Y, Dutta P, Davuluri R. Deep multi-omics integration by learning correlation-maximizing representation identifies prognostically stratified cancer subtypes. Bioinformatics advances. 2023;3(1):vbad075.

[9] Huynh KLA, Tyc KM, Matuck BF, Easter QT, Pratapa A, Kumar N V, et al. Spatial Deconvolution of Cell Types and Cell States at Scale Utilizing TACIT. bioRxiv. 2024 Jun;

[10] Minoura K, Abe K, Nam H, Nishikawa H, Shimamura T. A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. Cell reports methods. 2021 Sep;1(5):100071.

[11] Liu S, Wang H, Li S, Zhang C. Mixture-of-experts and semantic-guided network for brain tumor segmentation with missing MRI modalities. Med Biol Eng Comput. 2024 Oct;62(10):3179–91.

[12] Ji Y, Dutta P, Davuluri R. Deep multi-omics integration by learning correlation-maximizing representation identifies prognostically stratified cancer subtypes. Bioinformatics advances. 2023;3(1):vbad075.

[13] Athaya T, Ripan RC, Li X, Hu H. Multimodal deep learning approaches for single-cell multi-omics data integration. Brief Bioinform. 2023 Sep;24(5).

[14] Picard M, Scott-Boyer MP, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021;19:3735–46.

[15] Li X, Li MW, Zhang YN, Xu HM. [Common cancer genetic analysis methods and application study based on TCGA database]. Yi Chuan. 2019 Mar;41(3):234–42.

[16] Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review. Brief Bioinform. 2022 Mar;23(2).

[17] Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review. Brief Bioinform. 2022 Mar;23(2).

[18] Gong P, Cheng L, Zhang Z, Meng A, Li E, Chen J, et al. Multi-omics integration method based on attention deep learning network for biomedical data classification. Comput Methods Programs Biomed. 2023 Apr;231:107377.

[19] Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning. Brief Bioinform. 2022 Jan;23(1).

[20] Brombacher E, Hackenberg M, Kreutz C, Binder H, Treppner M. The performance of deep generative models for learning joint embeddings of single-cell multi-omics data. Front Mol Biosci. 2022;9:962644.

[21] Kang M, Ko E, Mersha TB. A roadmap for multi-omics data integration using deep learning. Brief Bioinform. 2022 Jan;23(1).

[22] Erfanian N, Heydari AA, Feriz AM, Iañez P, Derakhshani A, Ghasemigol M, et al. Deep learning applications in single-cell genomics and transcriptomics data analysis. Biomed Pharmacother. 2023 Sep;165:115077.

[23] Picard M, Scott-Boyer MP, Bodein A, Périn O, Droit A. Integration strategies of multi-omics data for machine learning analysis. Comput Struct Biotechnol J. 2021;19:3735–46.

[24] Wang C, Lye X, Kaalia R, Kumar P, Rajapakse JC. Deep learning and multi-omics approach to predict drug responses in cancer. BMC Bioinformatics. 2022 Nov;22(Suppl 10):632.

[25] Athaya T, Ripan RC, Li X, Hu H. Multimodal deep learning approaches for single-cell multi-omics data integration. Brief Bioinform. 2023 Sep;24(5).

[26] Zhang X, Xing Y, Sun K, Guo Y. OmiEmbed: A Unified Multi-Task Deep Learning Framework for Multi-Omics Data. Cancers (Basel). 2021 Jun;13(12).