Annals of Healthcare Systems Engineering

www.ahse.reapress.com

Ann. Healthc. Syst. Eng. Vol. 3, No. 3 (2026) 159–170.

Paper Type: Original Article

Tumor Type Classification from Somatic Mutations Using a Transformer-Based Model

Saeid Sanaei[1], Reza Javanmard Alitappeh[1,*]

[1] Faculty of Electrical and Computer Engineering, Mazandaran University of Science and Technology, Behshahr, Iran; saeid.sanaei22@gmail.com; rezajavanmard64@gmail.com.

Citation:

Received: 27 July 2025 *Revised: 15 October 2025* *Accepted: 10 January 2026*	Sanaei, S., & Javanmard Alitappeh, R. (2026). Tumor type classification from somatic mutations using a transformer-based model. *Annals of Healthcare Systems Engineering, 3*(3), 159-170.

Abstract

Somatic mutation profiles are sparse and variable in length, which makes tumor-type classification difficult with standard fixed-vector inputs. We propose a transformer-based model that represents each The Cancer Genome Atlas (TCGA) sample as a sequence of mutation tokens combining gene, genomic-position, substitution, sequence-context, variant-type, Variant Allele Fraction (VAF), read-depth, hotspot, and consequence features. Categorical fields were encoded by stable hashing, and numerical fields were normalized using only the training split of each fold. After filtering tumor types with at least 180 samples, 8,910 samples from 20 classes were evaluated. In stratified eight-fold cross-validation, the model achieved 64.4% accuracy, 65.4% weighted accuracy, 95.8% one-vs-rest ROC-AUC, 86.7% Top-3 accuracy, and 93.1% Top-5 accuracy. These results support sequence-based mutation modeling for ranked tumor-type prediction.

Keywords: Tumor type classification, Somatic mutations, Transformer model, Deep learning, TCGA, Precision oncology.

1 | Introduction

Sequencing-based cancer studies routinely produce mutation annotation files that describe many somatic variants for each sample. These records contain useful clues about tumor type, but they are not naturally organized as a fixed-length feature vector. One patient may have only a small number of reported variants, while another may have hundreds. In addition, most individual variants occur rarely across the cohort. The modeling problem is therefore to summarize a variable-length mutation profile into a sample-level representation while preserving the information carried by individual mutation records [1], [2].

Previous work has shown that tumor type can be inferred not only from well-known driver genes but also from broader passenger mutation patterns [3]. This finding is useful because it implies that tissue-of-origin

Corresponding Author: Saeid.sanaei22@gmail.com
https://doi.org/10.22105/ahse.v3i3.67

information is distributed across the mutation profile. At the same time, mutation profiles can be affected by cohort composition, sequencing depth, preprocessing choices, and class imbalance. For this reason, a reliable evaluation must avoid leakage between training and validation data and should report more than a single accuracy value.

Deep learning methods have been used to address the weak and distributed nature of mutation-derived signals. Mutation-Attention (MuAt) learns representations of mutation contexts and aggregates them for tumor typing and subtyping [2]. Multiple-instance learning is another natural choice for this setting, because the label is assigned to the sample, whereas the observations are individual mutations [1]. In both cases, the model must learn how mutation-level information should be pooled into a patient-level prediction.

Other studies have used vector-space representations of genomic alterations with classical machine-learning models [4], reduced neural networks for lung cancer histology [5], mutational signatures for primary-metastatic discrimination [6], and genome-wide mutation features for cancer of unknown primary [7], [8]. These studies support the general idea that mutation-derived features are informative, while also showing that no single representation is sufficient for all clinical and pan-cancer settings.

The clinical motivation is clearest when the tissue of origin is uncertain or when molecular profiling is the main available evidence. Related approaches have used targeted sequencing, nationwide genomic profiling, sparse liquid-biopsy mutation profiles, cfDNA methylation, and transcriptome-based assays to prioritize candidate tissues of origin [7–13]. Mutation-based models are a practical part of this broader effort because somatic variants are already produced by many cancer sequencing workflows.

Mutation profiles are also used for related prediction tasks. Microsatellite instability has been predicted from NGS data using multiple-instance learning and tabular methods such as NGBoost-based models [14], [15]. These tasks share several technical issues with tumor typing: Variable input length, high-dimensional categorical fields, class imbalance, and the risk that preprocessing statistics leak information from the validation set.

In this study, tumor-type classification from The Cancer Genome Atlas (TCGA) somatic mutations is formulated as a sequence modeling problem. Each mutation is converted into a token containing categorical and numerical descriptors, and a transformer encoder is used to model relations among the tokens within the same sample. A learnable (CLS) token is used to obtain the sample-level representation for 20-class tumor-type prediction.

The work makes three practical contributions. First, it keeps the mutation-level structure of the data instead of immediately reducing each sample to a handcrafted vector. Second, it uses stable hashed encoding for categorical variables and fold-specific normalization for numerical variables to reduce avoidable leakage. Third, it reports cross-validated accuracy, weighted accuracy, Top-k accuracy, ROC-AUC, and confusion-matrix behavior in a consistent experimental setting.

2 | Materials and Methods

2.1 | Data Source and Study Design

The data were obtained from TCGA. For each sample, the tumor label is defined at the sample level, while the input information consists of a set of somatic mutation records. This structure differs from ordinary tabular classification, because both the number of mutations and the content of the mutation list vary from sample to sample [1], [2].

Mutation records followed a Mutation Annotation Format (MAF)-like structure. The fields used in this work included gene symbol, chromosome, genomic position, reference and alternate alleles, substitution type, local sequence context, variant consequence class, hotspot indicator, and allele read counts when available. These fields form the mutation-level input summarized in *Fig. 1*.

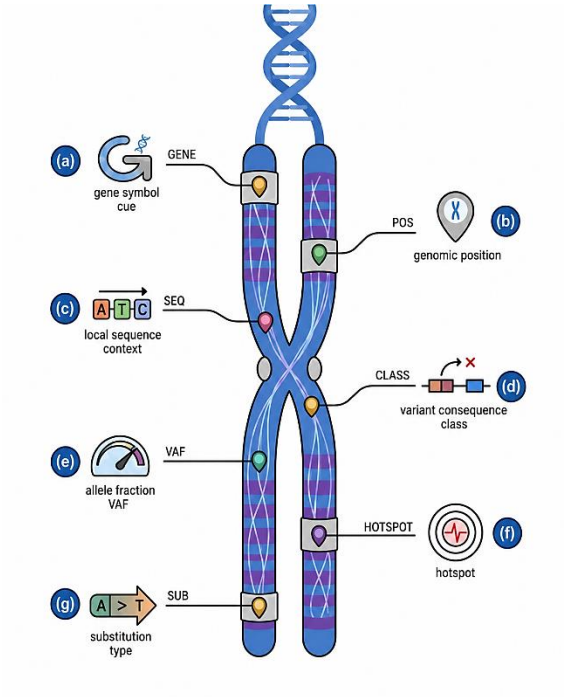

Fig. 1. Mutation-level input features used in the proposed model: (a) Gene symbol cue, (b) Genomic position, (c) Local sequence context, (d) Variant consequence class, (e) Variant Allele Fraction (VAF), (f) Hotspot indicator, and (g) Substitution type.

2.2 | Preprocessing and Class Definition

Sample identifiers were aligned with mutation records so that each mutation belonged to one sample only. Obvious duplicates and inconsistent records were removed. Tumor labels were derived from TCGA cancer-type codes. Rectal cancer samples were merged with colon cancer samples to obtain a single colorectal class, because the two groups are closely related and separating them would produce smaller classes.

Tumor types with fewer than 180 samples were excluded. After this filtering step, the dataset contained 8,910 samples from 20 tumor classes. The retained classes included BLCA, CESC, COAD, LUAD, LUSC, and other TCGA cancer types. Labels were represented as one-hot vectors during training and evaluation.

When tumor reference and alternate read counts were available, Variant Allele Fraction (VAF) was calculated and used as a numerical feature. VAF is a simple normalized measure of the fraction of reads supporting the alternate allele:

$$\mathrm{VAF} = \frac{\mathrm{alt\_count}}{\mathrm{ref\_count}+\mathrm{alt\_count}}. \quad (1)$$

In *Eq. (1)*, alt_count denotes the number of reads supporting the alternate allele and ref_count denotes the number of reads supporting the reference allele in the tumor sample. This value ranges from 0 to 1 and was used as a numerical feature together with the other mutation-level features.

2.3 | Mutation-Token Construction

Each sample was represented as a sequence of mutation tokens. A token stores the descriptors of one mutation, and the full sequence represents the mutation profile of the sample. This design leaves the aggregation step to the model rather than imposing a fixed handcrafted summary at the beginning of the pipeline.

Since mutation counts differ between samples, all input sequences were brought to a fixed maximum length. Let M_i be the raw number of mutations observed in sample i. The effective length L_i was defined as:

$$L_i = \min(M_i, 512). \quad (2)$$

Samples with more than 512 mutations were truncated using a fixed and reproducible ordering. Samples with fewer than 512 mutations were padded with neutral values. A binary mask was used to separate observed mutation tokens from padding positions:

$$m_{i,t} = \begin{cases} 1, & t \le L_i, \\ 0, & t > L_i. \end{cases} \quad (3)$$

The mask was applied during self-attention so that padded positions did not contribute to contextual token interactions. This is important because differences in mutation count should not be confused with artificial patterns introduced by padding.

2.4 | Feature Encoding and Leakage Control

Each token combined categorical and numerical descriptors. The categorical descriptors included gene identifier, chromosome, substitution type, sequence context, variant type, and variant consequence group. The numerical descriptors included VAF, read depth, genomic position, and hotspot indicator. High-cardinality categorical fields were mapped into fixed buckets using stable hashing, which avoided building a vocabulary from the full dataset.

Numerical features were normalized inside each cross-validation fold. For every fold, the normalization parameters were estimated from the training split and then applied to the held-out split. This step was used to prevent validation data from influencing preprocessing statistics.

After encoding, categorical features were passed through learnable embedding layers and numerical features were projected into the same dimensional space. The final mutation-token representation was computed as

$$x_{i,t} = \mathrm{Proj}_{\mathrm{num}}(r_{i,t}) + \mathrm{Emb}_{\mathrm{cat}}(z_{i,t}). \quad (4)$$

In *Eq. (4)*, $x_{i,t}$ is the token representation of mutation t in sample i. The vector $z_{i,t}$ contains the categorical identifiers, $\mathrm{Emb}_{\mathrm{cat}}$ denotes the categorical embedding operation, $r_{i,t}$ contains the numerical features, and $\mathrm{Proj}_{\mathrm{num}}$ maps numerical inputs to the token dimension. This stage produces an input tensor of size N x 512 x d and a mask of size N x 512, where N is the number of samples and d is the token dimension.

Categorical embeddings and numerical projections were combined by additive fusion, followed by layer normalization and dropout. The overall workflow from mutation records to tumor-type probabilities is summarized in *Fig. 2*.

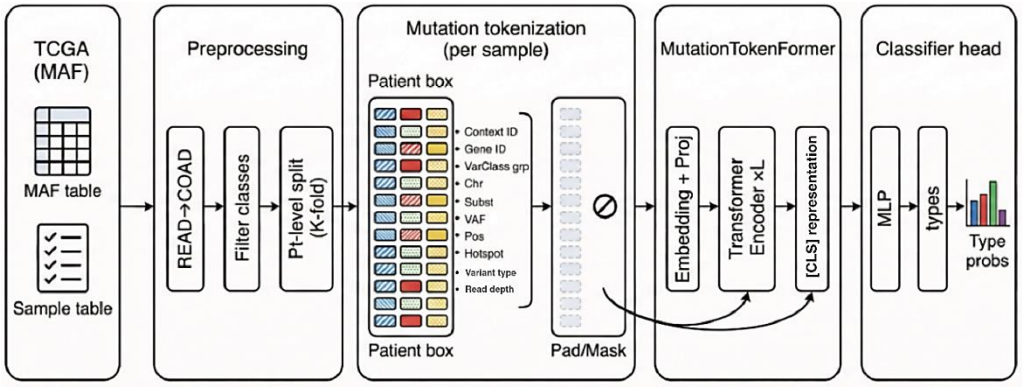

Fig. 2. Overall workflow of the proposed MutationTokenFormer model.

2.5 | Transformer Encoder and Classification Head

The overall workflow in *Fig. 2* summarizes the complete processing path, while *Fig. 3* provides a closer view of the sequence assembly and input-encoding steps. In each sample, mutation records are converted into a fixed-length sequence, an attention mask is created, categorical and numerical features are fused, and a (CLS) token is added before transformer encoding.

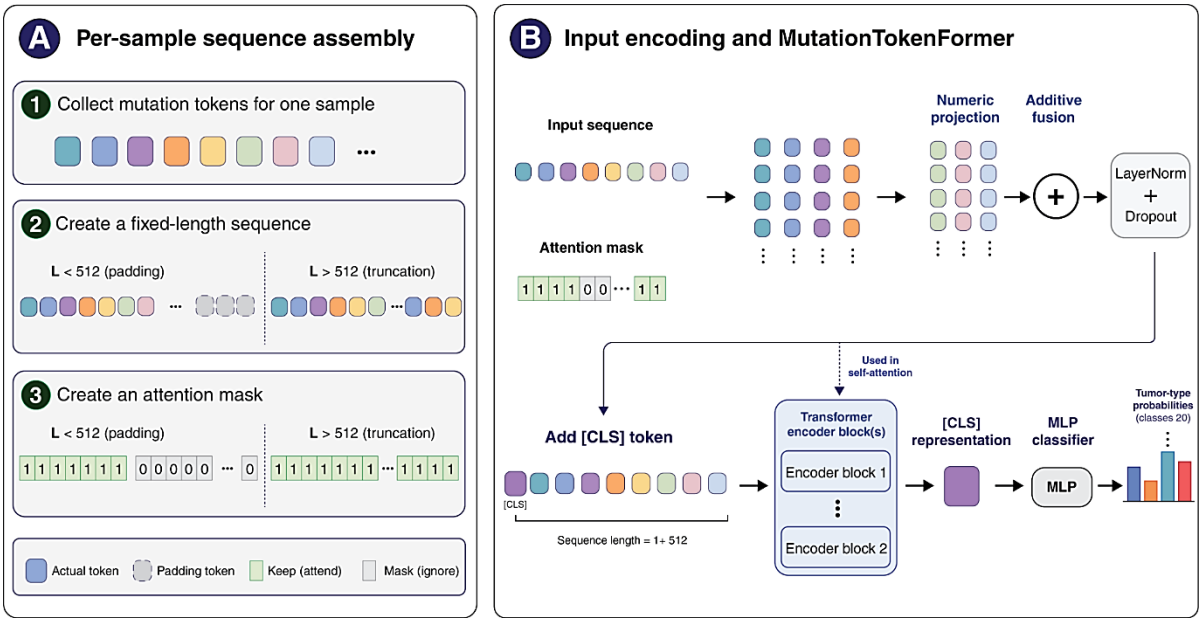

Fig. 3. Detailed sequence assembly and input encoding in MutationTokenFormer.

The encoded sequence was then passed to the transformer encoder. The attention mask was used in self-attention to ignore padded positions, while the (CLS)] token summarized mutation-level interactions for sample-level classification.

After the transformer layers, the contextualized (CLS) token was used as the sample-level representation:

$$h_i = z_{i,CLS.} \quad (5)$$

In *Eq. (5)*, $z_{i,CLS}$ denotes the transformer output corresponding to the (CLS) token for sample i, and h_i is the final sample-level representation. A multilayer perceptron then produced one logit for each of the 20 tumor classes, and class probabilities were obtained using the softmax function.

2.6 | Evaluation Protocol

Performance was estimated with stratified eight-fold cross-validation. In each fold, all preprocessing steps that depended on data statistics were fitted on the training split only. Predictions for the held-out split were saved, and the out-of-fold predictions from all folds were pooled for the final analysis.

The main metrics were overall accuracy, weighted accuracy, and one-vs-rest ROC-AUC. Top-1, Top-3, and Top-5 accuracies were also reported to evaluate whether the correct tumor type was ranked among the most likely candidate classes. ROC curves and a row-normalized confusion matrix were used for class-level inspection.

3 | Results

3.1 | Input-Sequence Characteristics

The retained cohort contained 8,910 samples across 20 tumor types. Mutation-token counts were highly right-skewed: The median number of mutation tokens per sample was 101.0, whereas the mean was 349.2. Overall, 88.8% of samples contained fewer than the fixed sequence length of 512 tokens and therefore required padding, while 11.2% exceeded this length and were truncated.

Fig. 4 summarizes mutation-token counts and the fraction of samples affected by the 512 token limit across tumor types. The largest percentages of truncated samples were observed for SKCM and UCEC, whereas THCA and KIRP were entirely below the 512 token limit.

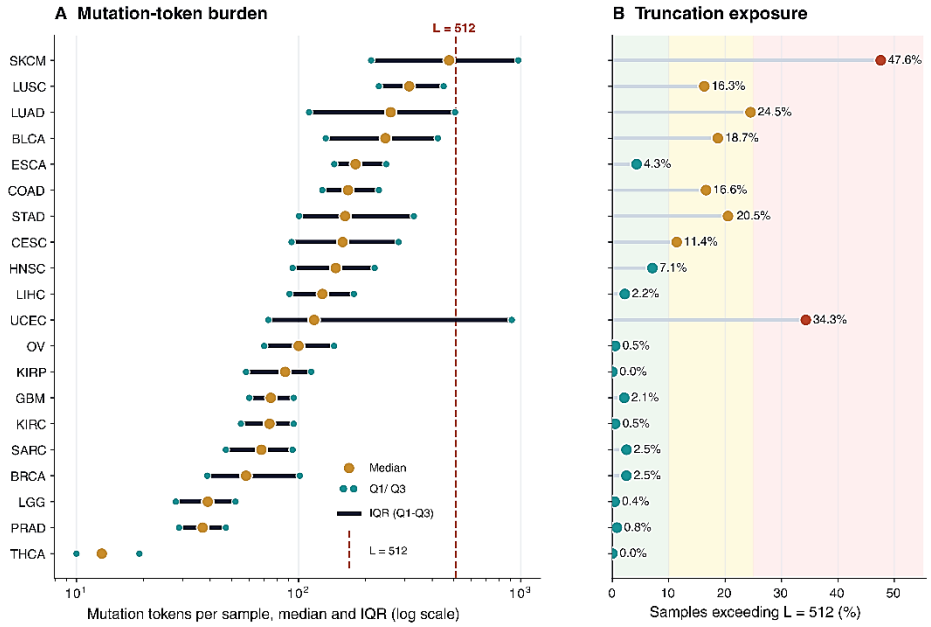

Fig. 4. Input-sequence characteristics across tumor types.

3.2 | Overall Classification Performance

Out-of-fold predictions from stratified eight-fold cross-validation were pooled before calculating the final metrics. As shown in Table 1, the proposed model achieved 64.4% accuracy, 65.4% weighted accuracy, and 95.8% one-vs-rest ROC-AUC.

Compared with the ATGC reference result, the proposed model improved both accuracy and weighted accuracy while maintaining a high ROC-AUC. This indicates that the mutation-token representation improved the final class assignment without reducing the ranking quality of the model.

Table 1. Performance comparison between the proposed method and ATGC.

Method	Accuracy	Weighted Accuracy	ROC-AUC
ATGC [1]	58.0	59.8	95.2
Proposed method	64.4	65.4	95.8

3.3 | Top-k Classification Performance

Top-k metrics were used to evaluate whether the correct tumor type appeared among the highest-probability candidate classes. The proposed model achieved 64.4% Top-1 accuracy, 86.7% Top-3 accuracy, and 93.1% Top-5 accuracy, as shown in *Table 2*.

Table 2. Top-k performance comparison with MuAt and DNN.

Method	Top-1	Top-3	Top-5
MuAt [2]	64.1	83.5	90.6
DNN [2]	62.8	82.9	91.8
Proposed method	64.4	86.7	93.1

Among the 3,168 samples that were incorrect at Top-1, 62.6% were recovered within the Top-3 predictions and 80.5% were recovered within the Top-5 predictions (*Table 3*). This supports the use of ranked candidate lists rather than a single forced label in mutation-only tumor typing.

Table 3. Recovery of Top-1 errors within Top-3 and Top-5 predictions.

Condition	Number of Samples	Percentage of Top-1 Errors (%)
Top-1 errors	3168	100
Recovered within Top-3	1982	62.6
Recovered within Top-5	2549	80.5
Not recovered within Top-5	619	19.5

3.4 | ROC Analysis

Multi-class ROC curves were computed in a one-vs-rest setting. *Fig. 5* shows the micro-average and macro-average ROC curves. The micro-average ROC-AUC was 0.959 and the macro-average ROC-AUC was 0.958.

The closeness of the micro- and macro-average AUC values suggests that discrimination was not driven only by the largest tumor classes. ROC-AUC also captures ranking quality, whereas accuracy depends on the final selected class.

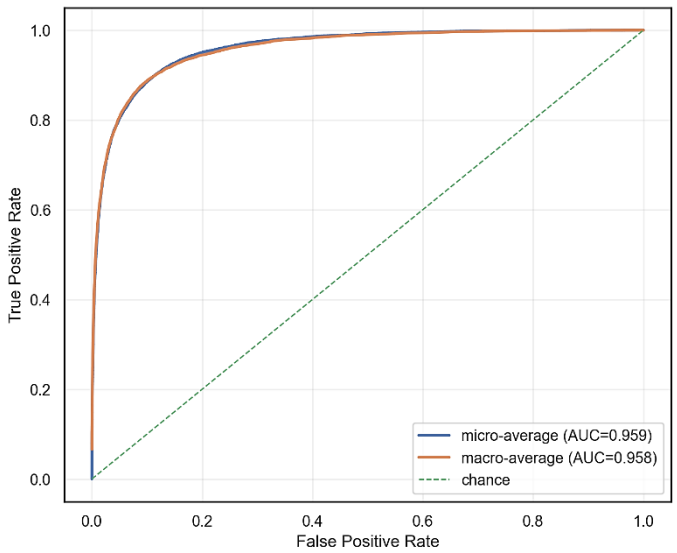

Fig. 5. Multi-class one-vs-rest ROC curves for tumor-type prediction.

3.5 | Class-Wise Performance Analysis

Class-wise precision, recall, and F1-score were calculated from the pooled out-of-fold predictions. *Table 4* shows that performance varied across tumor types. The highest F1-score was obtained for SKCM (90.2%), while lower values were observed for HNSC (46.8%) and PRAD (49.4%).

Table 4. Class-wise performance of the proposed model.

Tumor type	Support	Prec	Rec	F1
BLCA	411	56.4	68.4	61.8
BRCA	1020	60.8	50.4	55.1
CESC	289	59.8	68.5	63.9
COAD	517	67.8	81.6	74.1
ESCA	184	61.6	62	61.8
GBM	390	63.1	62.3	62.7
HNSC	507	47.3	46.4	46.8
KIRC	369	71.9	74.8	73.3
KIRP	281	74	61.9	67.4
LGG	510	71.9	70.8	71.3
LIHC	363	72.7	74.1	73.4
LUAD	515	65.7	59	62.2
LUSC	485	58.8	68	63.1
OV	411	58.1	74.7	65.4
PRAD	495	49.9	48.9	49.4
SARC	236	57.6	57.6	57.6
SKCM	466	94.2	86.5	90.2
STAD	439	62.6	67.9	65.1
THCA	492	79.6	60.2	68.5
UCEC	530	66.9	64	65.4

The ranked F1-score profile in *Fig. 6* provides a compact visual summary of this class-level variation. The results show that the model separated several tumor types reliably, but mutation-only information remained less distinctive for some classes.

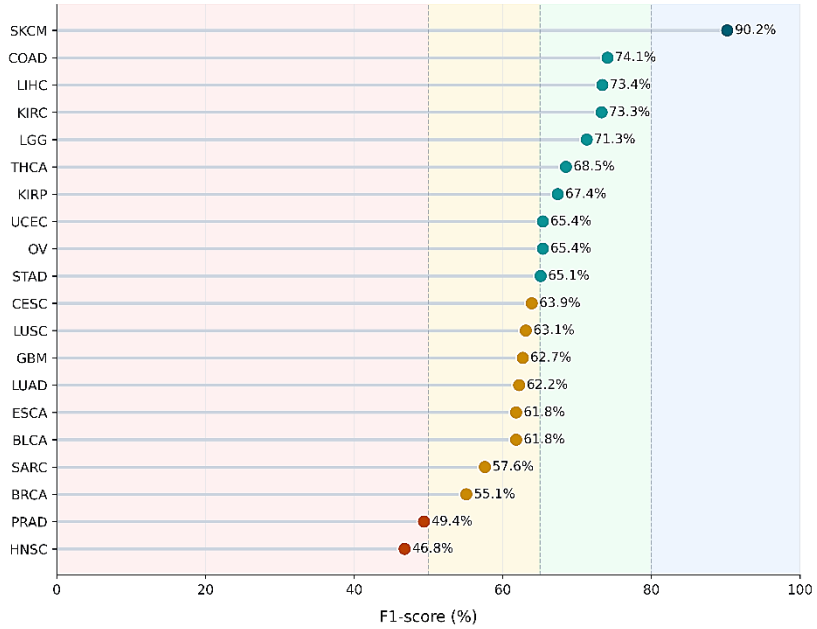

Fig. 6. Class-wise F1-score across tumor types.

3.6 | Confusion-Matrix and Error-Pattern Analysis

A row-normalized confusion matrix was used to inspect error patterns. As shown in *Fig. 7*, most classes had visible diagonal dominance, but off-diagonal errors were concentrated in specific tumor-type pairs rather than being uniformly distributed.

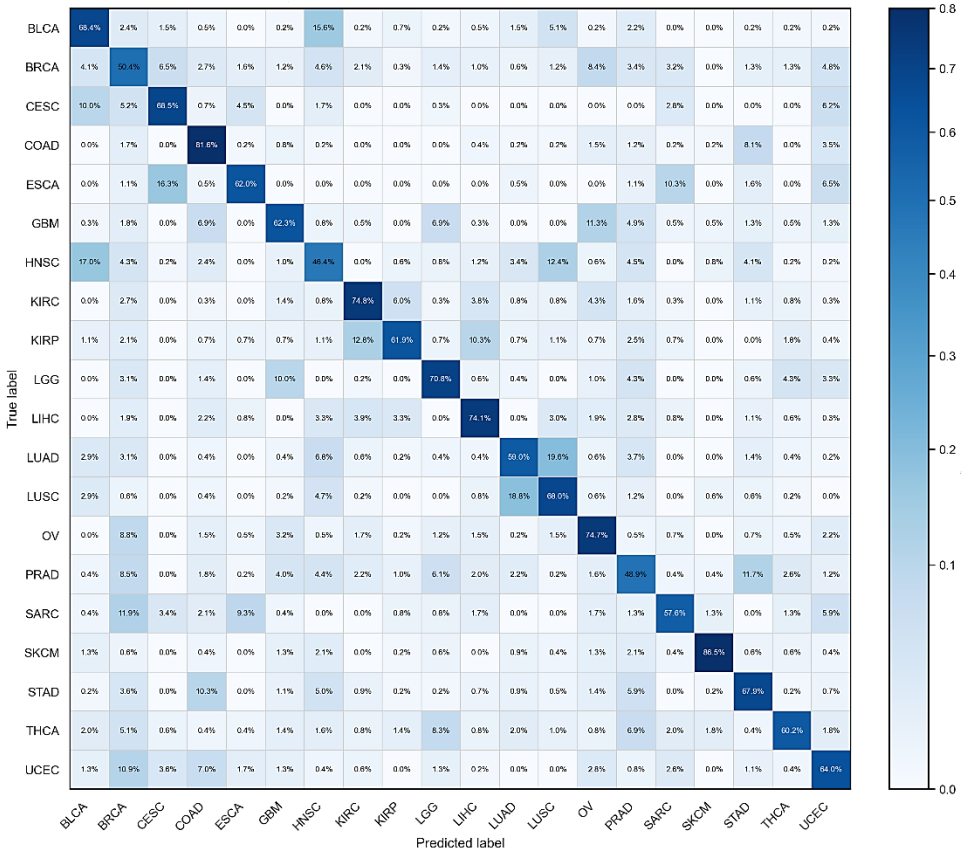

Fig. 7. Row-normalized confusion matrix of tumor-type predictions.

The most frequent confusion pairs are listed in *Table 5*. LUAD-to-LUSC and LUSC-to-LUAD were the two largest error patterns, followed by HNSC-to-BLCA and BRCA-to-OV. These patterns are useful for identifying tumor groups that may require additional molecular features or a hierarchical decision strategy.

Table 5. Most frequent tumor-type confusions.

True Tumor Type	Predicted Tumor Type	Misclassified Samples	Percentage of True Class (%)
LUAD	LUSC	101	19.6
LUSC	LUAD	91	18.8
HNSC	BLCA	86	17
BRCA	OV	86	8.4
BRCA	CESC	66	6.5
BLCA	HNSC	64	15.6
HNSC	LUSC	63	12.4
PRAD	STAD	58	11.7
UCEC	BRCA	58	10.9
LGG	GBM	51	10
BRCA	UCEC	49	4.8
BRCA	HNSC	47	4.6
STAD	COAD	45	10.3
GBM	OV	44	11.3
PRAD	BRCA	42	8.5

4 | Discussion

The results show that keeping mutation records as tokens is a reasonable strategy for TCGA tumor-type classification. Instead of compressing all variants into a fixed handcrafted vector before modeling, the transformer receives a structured list of mutation-level observations and learns how to combine them at the sample level through a contextualized (CLS) representation.

The input-sequence analysis clarifies why padding, truncation, and masking were needed. The cohort had a median of 101 mutation tokens per sample, but the mean was 349.2 because a smaller subset of highly mutated samples created a long right tail. Most samples required padding, while 11.2% exceeded the 512 token limit and were truncated. This shows that the model had to handle both sparse and highly mutated profiles within the same evaluation setting.

The improvement over ATGC in both accuracy and weighted accuracy suggests that the token-based representation captured information that was useful for final class assignment. The high Top-3 and Top-5 values are also important. In practice, a ranked list of candidate tumor types may be more useful than a single forced label, especially when mutation profiles are incomplete or when biologically related tumors share similar alterations.

The class-wise analysis showed substantial variation among tumor types. SKCM achieved the strongest F1-score, whereas HNSC and PRAD were more difficult. These differences are consistent with the idea that mutation-only signals are not equally distinctive across all cancer types. Some classes may require additional information, such as copy-number alteration, methylation, expression, mutational signatures, or clinical variables, to improve separation.

The confusion matrix and the ranked confusion table provide a more detailed view of the remaining errors. The largest error patterns occurred between LUAD and LUSC and between several biologically or clinically related tumor groups. Such overlap is expected in a pan-cancer mutation-only setting and may be reduced by using hierarchical classifiers or multimodal molecular inputs.

The preprocessing choices were selected to keep the evaluation conservative. Stable hashing reduced the need for a vocabulary built from all samples, and fold-specific normalization ensured that validation statistics were not used during training. These details are important in mutation-profile studies, where leakage can easily inflate performance estimates.

The findings are consistent with previous mutation-based tumor-typing studies. Multiple-instance learning treats each sample as a bag of mutation instances [1], while MuAt and related attention-based models learn representations from mutation context [2]. Other studies based on targeted sequencing, vector-space transformations, and large genomic profiling cohorts also show that mutation-derived features can support tumor classification [4], [9], [10].

The study has limitations. First, the evaluation was limited to TCGA and did not include an independent external cohort. TCGA is useful for method development, but it does not cover all sequencing platforms, sample qualities, or clinical populations. Second, the maximum sequence length was fixed at 512 mutations, which may truncate information in highly mutated tumors. Third, the model used mutation-derived features only and did not integrate other molecular or clinical information.

A further limitation is the absence of a detailed ablation study. Future experiments should measure the separate contribution of gene identity, sequence context, variant type, consequence class, VAF, read depth, hotspot status, sequence ordering, (CLS)-based sequence summarization, and class-imbalance handling. Such analysis would show which parts of the token representation are most responsible for performance.

Future work should also test the model on independent multi-center data and evaluate calibration if the outputs are used as probabilities. A hierarchical design could first predict broad tissue groups and then

distinguish closely related subtypes. Combining mutation tokens with other omics layers may reduce the remaining confusion among biologically similar cancers.

5 | Conclusion

This work presented a transformer-based model for tumor-type classification from somatic mutations. Each sample was represented as a sequence of mutation tokens, and a contextualized (CLS) token was used as the sample-level representation for classification. The evaluation used TCGA data after retaining tumor classes with at least 180 samples, resulting in 8,910 samples and 20 classes.

The model achieved 64.4% overall accuracy, 65.4% weighted accuracy, and 95.8% one-vs-rest ROC-AUC. It also achieved 86.7% Top-3 accuracy and 93.1% Top-5 accuracy. Class-wise and confusion-matrix analyses showed that the model performed strongly for several tumor types but still confused some related classes. Future work should focus on external validation, ablation studies, calibration, and integration of additional molecular features.

Author's Contributions

Conceptualization, S.S. and R.J.A.T.; methodology, S.S.; software, S.S.; validation, S.S. and R.J.A.T.; formal analysis, S.S.; investigation, S.S.; data curation, S.S.; writing-original draft, S.S.; writing-review and editing, S.S. and R.J.A.T.; visualization, S.S.; supervision, R.J.A.T. All authors have read and agreed to the final version of the manuscript.

Funding

The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data Availability

In this study, publicly available datasets were analyzed. These data can be found in TCGA through the Genomic Data Commons (GDC) Data Portal. Processed data tables and analysis scripts are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare no conflict of interest. Funders played no role in the design of the study, in the collection, analysis, or interpretation of data, in the writing of the manuscript, or in the decision to publish the results.

Ethics Approval and Consent to Participate

This study used publicly available, de-identified genomic data and did not involve new recruitment of human participants or new collection of biospecimens. Therefore, no additional institutional ethics approval was required for the analyses reported here.

References

[1] Anaya, J., Sidhom, J. W., Mahmood, F., & Baras, A. S. (2024). Multiple-instance learning of somatic mutations for the classification of tumour type and the prediction of microsatellite status. *Nature Biomedical Engineering*, *8*(1), 57–67. https://doi.org/10.1038/s41551-023-01120-3

[2] Sanjaya, P., Maljanen, K., Katainen, R., Waszak, S. M., Ambrose, J. C., Arumugam, P., … ., & Consortium, G. E. R. (2023). Mutation-Attention (MuAt): Deep representation learning of somatic mutations for tumour typing and subtyping. *Genome Medicine*, *15*(1), 47. https://doi.org/10.1186/s13073-023-01204-4

[3] Jiao, W., Atwal, G., Polak, P., Karlic, R., Cuppen, E., Al-Shahrour, F., … ., & Consortium, P. (2020). A deep learning system accurately classifies primary and metastatic cancers using passenger mutation patterns. *Nature Communications*, *11*(1), 728. https://doi.org/10.1038/s41467-019-13825-8

[4] Zelli, V., Manno, A., Compagnoni, C., Ibraheem, R. O., Zazzeroni, F., Alesse, E., … ., & Tessitore, A. (2023). Classification of tumor types using XGBoost machine learning model: A vector space transformation of genomic alterations. *Journal of Translational Medicine*, *21*(1), 836. https://doi.org/10.1186/s12967-023-04720-4

[5] Kobayashi, K., Bolatkan, A., Shiina, S., & Hamamoto, R. (2020). Fully-connected neural networks with reduced parameterization for predicting histological types of lung cancer from somatic mutations. *Biomolecules*, *10*(9), 1–16. https://doi.org/10.3390/biom10091249

[6] Zheng, W., Pu, M., Li, X., Du, Z., Jin, S., Li, X., … ., & Zhang, Y. (2023). Deep learning model accurately classifies metastatic tumors from primary tumors based on mutational signatures. *Scientific Reports*, *13*(1), 8752. https://doi.org/10.1038/s41598-023-35842-w

[7] Nguyen, L., Van Hoeck, A., & Cuppen, E. (2022). Machine learning-based tissue of origin classification for cancer of unknown primary diagnostics using genome-wide mutation features. *Nature communications*, *13*(1), 4013. https://doi.org/10.1038/s41467-022-31666-w

[8] Moon, I., LoPiccolo, J., Baca, S. C., Sholl, L. M., Kehl, K. L., Hassett, M. J., … ., & Gusev, A. (2023). Machine learning for genetics-based classification and treatment response prediction in cancer of unknown primary. *Nature Medicine*, *29*(8), 2057–2067. https://doi.org/10.1038/s41591-023-02482-6

[9] Darmofal, M., Suman, S., Atwal, G., Toomey, M., Chen, J. F., Chang, J. C., … ., & Morris, Q. (2024). Deep-learning model for tumor-type prediction using targeted clinical genomic sequencing data. *Cancer Discovery*, *14*(6), 1064–1081. https://doi.org/10.1158/2159-8290.CD-23-0996

[10] Huang, Y., Pfeiffer, S. M., & Zhang, Q. (2023). Primary tumor type prediction based on US nationwide genomic profiling data in 13,522 patients. *Computational and Structural Biotechnology Journal*, *21*, 3865–3874. https://doi.org/10.1016/j.csbj.2023.07.036

[11] Danyi, A., Jager, M., & de Ridder, J. (2021). Cancer type classification in liquid biopsies based on sparse mutational profiles enabled through data augmentation and integration. *Life*, *12*(1), 1. https://doi.org/10.3390/life12010001

[12] Conway, A. M., Pearce, S. P., Clipson, A., Hill, S. M., Chemi, F., Slane-Tan, D., … ., & Rothwell, D. G. (2024). A cfDNA methylation-based tissue-of-origin classifier for cancers of unknown primary. *Nature Communications*, *15*(1), 3292. https://doi.org/10.1038/s41467-024-47195-7

[13] Michuda, J., Breschi, A., Kapilivsky, J., Manghnani, K., McCarter, C., Hockenberry, A. J., … ., & Taxter, T. (2023). Validation of a transcriptome-based assay for classifying cancers of unknown primary origin. *Molecular Diagnosis & Therapy*, *27*(4), 499–511. https://doi.org/10.1007/s40291-023-00650-5

[14] Ziegler, J., Hechtman, J. F., Rana, S., Ptashkin, R. N., Jayakumaran, G., Middha, S., … ., & Zehir, A. (2025). A deep multiple instance learning framework improves microsatellite instability detection from tumor next generation sequencing. *Nature Communications*, *16*(1), 136. https://doi.org/10.1038/s41467-024-54970-z

[15] Chen, J., Wang, M., Zhao, D., Li, F., Wu, H., Liu, Q., & Li, S. (2023). MSINGB: A novel computational method based on NGBoost for identifying microsatellite instability status from tumor mutation annotation data. *Interdisciplinary Sciences: Computational Life Sciences*, *15*(1), 100–110. https://doi.org/10.1007/s12539-022-00544-w