The Journal of Basic and Clinical Dentistry

Original Research

ProtBERT-Based Prediction of Functional TRPV Sequences for Oral Cancer Pain

Pradeep Kumar Yadalam[1*]

1. Department of Periodontics, Saveetha Dental College, SIMATS, Saveetha University, Chennai, Tamil Nadu, India.

***Corresponding Author:** Pradeep Kumar Yadalam, Department of Periodontics, Saveetha Dental College, SIMATS, Saveetha University, Chennai 600077 Tamil Nadu, India. **Email:** pradeepkumar.sdc@saveetha.com Contact: 9941912405.

Citation: Yadalam P, K. ProtBERT-Based Prediction of Functional TRPV Sequences for Oral Cancer Pain. J Basic Clin Dent, 2025;2(2), 1–14.

Received: 5th September

Revised: 18th October

Accepted: 30th November

Published: 10th December

Copyright: © 2025 by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

Abstract.

Background and Objective: Oral Cancer often leads to opioid tolerance, affecting pain-sensing neurons. Certain TRPV agonists can activate the channel transiently, followed by desensitization, whereas antagonists block ligand-induced activation, preventing channel opening. Cancer changes the peripheral and central nervous systems, affecting tumor interactions, nociceptors, and the immune system. Activation of these receptors triggers pain perception and the release of second messengers. This study uses a large language model to find and forecast functional TRPV sequences for oral cancer pain. **Materials and Methods:** TRPV proteins were obtained from UniProt; id -Q8NER1, Q9HBA0, Q9NQA5, Q9H1D0, and Q9Y5S1 were identified and quality-checked. Deepbio analyzed FASTA sequences. Large language models and sequence prediction techniques were used for prediction and classification, including Protbert (a pre-trained protein sequence model), BiLSTM, LSTMAttention, and TextRGNN (a residual graph neural network). **Results:** The study reveals that ProtBERT and BiLSTM are the best predictors of amino acid sequences for TRPV1-based oral cancer, offering a balance of sensitivity and specificity. The ProtBERT model effectively identifies positive and negative cases related to TRPV with an accuracy of 0.90, a sensitivity of 0.89, a specificity of 0.91, and an AUC of 0.937. The BiLSTM model, like ProtBERT, achieves high accuracy and sensitivity, while the LSTM Attention model is reliable at detecting true positives. The text RGNN model has an accuracy of 0.51. **Conclusion:** Machine learning models such as ProtBERT and BiLSTM can enhance the early detection and diagnosis of oral cancer pain, but challenges such as data quality and validation remain.

Keywords: ProtBERT, BiLSTM, natural language processing, TRPV, oral cancer, Pain.

1. Introduction

Oral cancer, [1–5] which is painful and resistant to analgesics, often causes opioid tolerance. Functional TRPV (Transient Receptor Potential Vanilloid) ion channels are either homo- or hetero-tetrameric, with four identical subunits. In afferent nociceptors, pain-sensing neurons, four TRPVs [6–8] (TRPV1, TRPV2, TRPV3, and TRPV4) operate as transducers of temperature and chemical stimuli. The cancer microenvironment, characterized by elevated proton levels, is a TRPV1 agonist. TRPV1, a transient receptor potential vanilloid subtype 1, undergoes sensitization by inflammatory molecules like bradykinin (BK). [9,10] Cancer causes changes in the peripheral and central nervous systems through G-protein-coupled receptors (GPCRs), affecting tumor interactions, nociceptors, and the immune system, leading to variability in phenotypic and genomic characteristics. Sensory afferent terminals contain various receptors, including TRPA1, TRPV1, TRPV4, and several TRPM members, all of which play roles in pain perception. [11,12]

This approach preserves crucial motor and somatosensory functions vital for daily activities and quality of life. Activation of these receptors [13] triggers the production of second messengers, including phosphatidylinositol-4,5-biphosphate (PIP2), inositol triphosphate (IP3), and diacylglycerol (DAG). This intricate signaling cascade highlights the multifaceted regulatory pathways that contribute to TRPV1 sensitization in response to inflammatory stimuli. [14,15]

Oral cancer patients show increased sensitivity to spicy foods due to changes in the trigeminal ganglion. One study found increased TRPV1 expression in these neurons, leading to a greater capsaicin response. ERK1/2 and TRPV4 are involved in dorsal root ganglion (DRG) pain pathways during tumor progression, with TRPV1 and TLR4 co-expressing and increasing after inoculation. [16,17] Antagonists alleviate cancer-induced Pain, suggesting potential therapeutic approaches for neuropathic pain management. Capsazepine blocks certain TRP channels; garlic and ginger activate specific channels; and substances like nitric oxide and hydrogen peroxide interact with different channels after tissue damage. One study shows that TRPV1 antagonists and agonists influence pain behavior and TRPV1 expression in animal cancer models. TRPV1 antagonists reduced pain behaviors in mice. Increased TRPV1 in dorsal root ganglion neurons and peripheral axons after tumor inoculation indicates its role in cancer pain mechanisms.[9,10,18] Modulation and activation of TRPV1 contribute to oral cancer pain. Therefore, understanding TRPV1's role in this pain is crucial.

TRPV channels, specifically TRPV1,[17,19] are crucial in the sensory pathway for pain perception and inflammation. Predicting TRPV sequences is crucial for targeting oral cancer pain, given their roles in pain sensation, inflammation, and interactions with pain pathways, thereby aiding drug targeting and biomarker development. TRPV1 sequences can be profiled in cancer patients using advanced sequencing technologies and compared against known databases to identify variants. Computational modeling and functional assays can help elucidate and predict TRPV1 sequences, which are crucial for targeting pain mechanisms in oral cancer. [2,19] Amino acid sequence prediction is crucial in bioinformatics, drug discovery, and synthetic biology. Accurate prediction helps in drug discovery, biomarker identification, synthetic biology, vaccine development, and personalized medicine. Large language model approaches for sequence prediction include BERT (Bidirectional Encoder Representations from Transformers), BiLSTM (Bidirectional Long Short-

Term Memory Networks), and TextR-GNN (Text Representation using Graph Neural Networks). BERT is a transformer-based model that understands the context and relationships of sequences, while BiLSTM is a recurrent neural network that captures long-range dependencies in sequential data. TextR-GNN leverages relational structures within data to capture complex interdependencies between amino acid sequences, allowing for richer representations that improve predictions. [20,21] These models are vital for biomedical research and therapeutic development, leading to innovations in drug discovery, personalized medicine, and health diagnostics. [20]

BERT is a transformer-based language model designed to capture bidirectional contextual information, which allows it to outperform many conventional large language models in sequence inference tasks. [22] Google developed BERT as a bidirectional masked-language model to learn from unlabeled text, including biological data. [23] In drug discovery, BERT's masked aspect is crucial, enabling inference of missing structures in sequences such as FASTA. [24,25]

BertAIP, [22,26] a new tool for predicting anti-inflammatory peptides from protein sequences, outperforms other methods due to its unique model and focus on key amino acids. It's effective in drug development and inflammatory disease research, highlighting important amino acids. Feature vectors are extracted from protein sequences, encoded using a Graph-BERT model, and processed through a fully connected softmax layer for interaction classification. ProteinBERT, [27,28] it is a deep language model that integrates language modeling with Gene Ontology annotation prediction. It achieves state-of-the-art performance on protein properties benchmarks and enables the rapid training of protein predictors with limited labeled data. A previous study introduced AMP-BERT, a deep learning model that classified peptides into antimicrobial and non-antimicrobial categories. This model outperformed existing methods and employed an attention mechanism to enhance feature analysis, thereby improving its effectiveness for drug development. [23,29] Protein sequence–based prediction models have rapidly evolved, with recent studies demonstrating improved accuracy for ion channel related functional characterization using transformer-based architectures and hybrid recurrent models. Incorporating these computational advances provides a stronger justification for analyzing TRPV variants, especially in the context of oral cancer related nociception. [30]

On the other hand, GPT (Generative Pre-trained Transformer), a unidirectional model from OpenAI, [22,25] relies on large language modeling datasets to predict the next word in a sequence. Predicting sequences is crucial in drug discovery because it allows researchers to identify potential efficacy and interactions of novel compounds. Sophisticated models such as BERT and ProTBoRT enhance our understanding of intricate sequences by precisely analyzing and predicting protein sequences. [23] As a result, predicting sequences accelerates the discovery process and increases the likelihood of identifying effective and safe drug candidates. This study uses ProtBERT-based modelling to identify and predict pain-related TRPV variants in oral cancer.

2. Materials and Methods

Using UniProt IDs Q8NER1, Q9HBA0, Q9NQA5, Q9H1D0, and Q9Y5S1, the TRPV proteins were downloaded with a 100% match; similar proteins at 90% and 50% similarity were also retrieved in FASTA format. These sequences were then identified and quality-checked. Variable lengths are managed by padding or truncating sequences to a uniform length. In addition, tokenization converts amino acid sequences into tokens for model input. FASTA sequences underwent analysis using the Deepbio tool. The study classifies functional TRPV1 sequences linked to oral cancer pain, while class 0 refers to non-functional TRPV sequences or those not linked to pain pathways. The functional and non-functional labels were generated computationally based on whether TRPV1 sequences were annotated as pain-associated or not in the dataset, without implying biological validation. Class balance was addressed through stratified splitting to preserve proportional representation of both classes during training and testing.

Deep Bio [31] is a one-stop web service for academics to develop deep learning architectures for any biological topic. DeepBIO analyzed, optimized, and visualized biological sequencing data using deep-learning algorithms. Deep Bio separated sequence-based datasets into training and test data. We randomly split each dataset into 80% for training and 20% for testing to tune hyperparameters and evaluate performance. Large language models and other algorithms for sequence prediction used were like Protbert (pre-trained model on protein sequences), [22] BiLSTM [27] (Bidirectional Long Short-Term Memory), LSTMAttention (Long Short-Term Memory (LSTM with attention), TextRGNN (residual graph neural network for TEXT classification). Other models, like LSTM attention and textRGNN, were used to compare and validate the LLM (Figure 1).

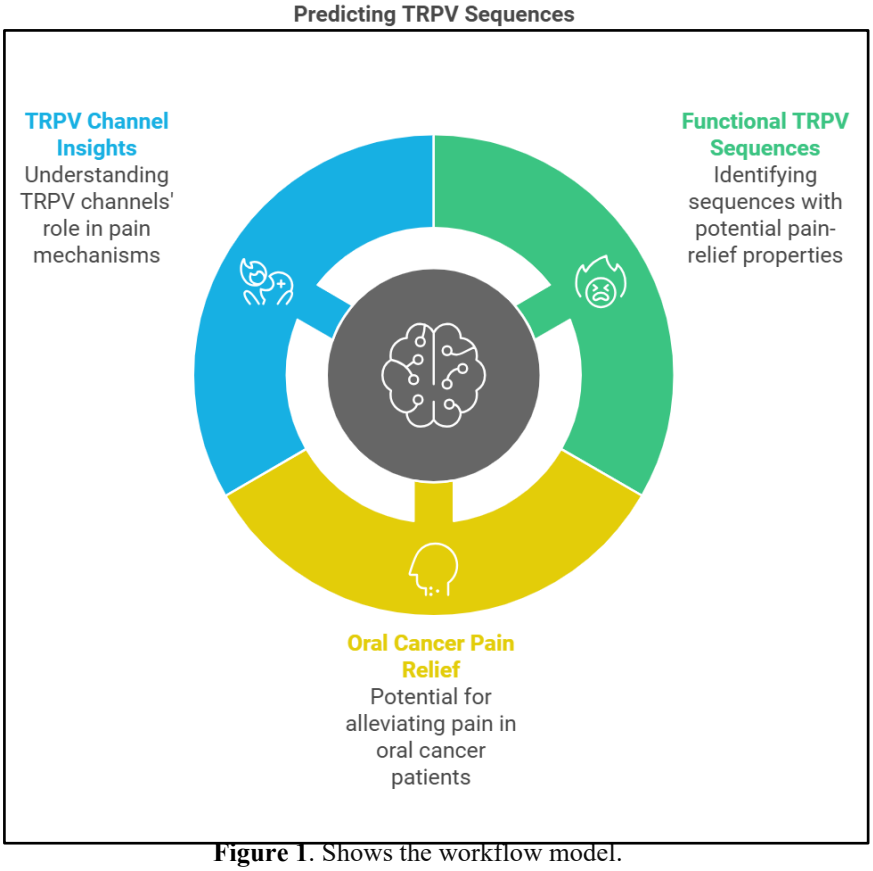

Figure 1. Shows the workflow model.

2. 1 Architecture and Hyperparameters

ProtBERT, [27] it is a pre-trained variant of the BERT model that uses a transformer architecture to capture contextual amino acid embeddings. It has 12 layers, 768 hidden units, 12 attention heads, 3072 intermediate units, the GeLU activation function, a 0.1 dropout rate, a maximum sequence length of 512, a varying learning rate, a batch size of 32-64, and Xavier initialization (Figure 2).

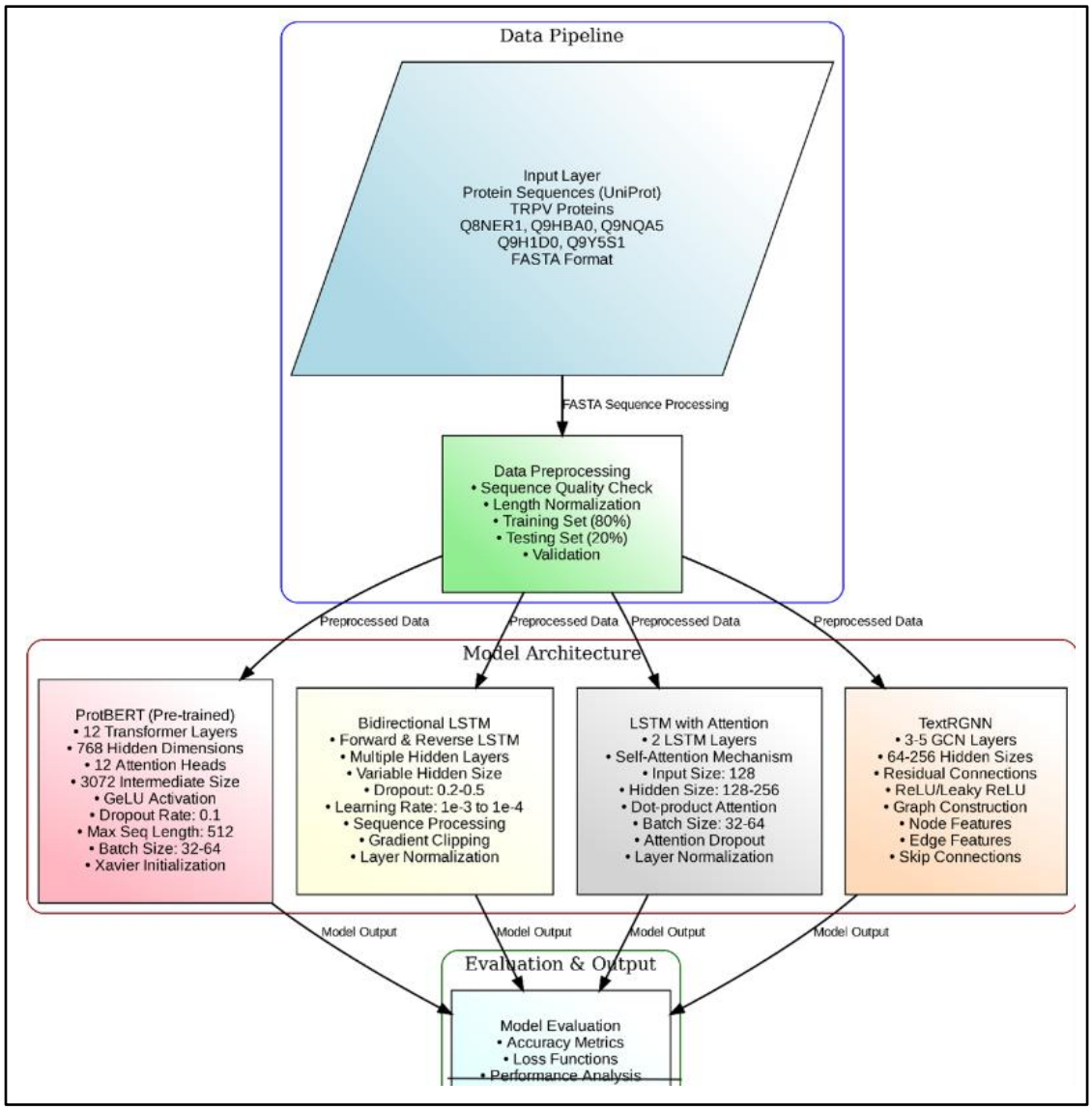

Figure 2. Shows the workflow schematics with the architecture

BiLSTM architecture uses two long short-term memory layers to process sequences in both forward and reverse directions. It has hyperparameters like input size, hidden size, number of layers, bidirectionality, dropout rate, learning rate, batch size, and sequence length. The model combines outputs from both forward and backward LSTMs. The hyperparameters include input size, hidden size, number of layers, bidirectional outputs, dropout rate, learning rate, batch size, and sequence length. They range from 1e-3 to 1e-4, with a learning rate of 1e-4.

The LSTMAttention Architecture combines an LSTM with an attention mechanism, enabling it to focus on specific parts of the input sequence for better performance on tasks that require contextual understanding. It has hyperparameters like input size, hidden size, number of LSTM layers, attention mechanism, dropout rate, learning rate, sequence length, and batch size. The input size of feature vectors is typically 128, with hidden sizes ranging from 128 to 256 based on task complexity. There are 2 LSTM layers, usually using dot-product or scaled dot-product attention. The dropout rate is 0.2 to 0.5, the learning rate is 1e-3 to 1e-4, the sequence length is flexible, and the batch size is 32 to 64.

TextRGNN is a residual graph neural network designed for text classification tasks. It models relationships between elements while preserving important features through residual connections. The graph construction method, based on co-occurrences, involves nodes representing words or characters and edges representing relationships between them. The text describes a graph convolutional model with 3-5 layers, hidden sizes of 64-256, residual connections, ReLU or Leaky ReLU activation functions, dropout rate, learning rate, batch size, and a flexible maximum sequence length.

3. Results

The model's accuracy, sensitivity, and specificity are key metrics in cancer pain detection. High accuracy indicates correct classification of positives and negatives. Sensitivity measures correctly identified positives; specificity measures correctly identified negatives. AUC summarizes performance across thresholds, with higher values indicating better ability. The ProtBERT model, with an accuracy of 0.90, sensitivity of 0.89, specificity of 0.91, and an AUC of 0.937, effectively identifies both positive and negative TRPV1 cases. The BiLSTM model, similar to ProtBERT, achieves high accuracy and sensitivity, with a specificity of 0.89 and an AUC of 0.945, indicating its reliability in detecting true positives. The LSTM Attention model is a strong predictive model, achieving 0.89 in accuracy, sensitivity, and specificity, and an AUC of 0.931. The text RGNN has an accuracy of 0.51, a sensitivity of 0.04, a specificity of 0.98, and an AUC of 0.657. The Text RGNN, a deep neural network, has a low accuracy of 0.51, indicating it identifies only about half of cases, compared to other models with higher accuracy. Its low sensitivity of 0.04, indicating it identifies only 4% of actual positives, is crucial in clinical contexts where detecting true positives can have significant negative consequences. However, its high specificity of 0.98 is not useful if the model fails to detect genuine positive cases. The AUC of 0.657 indicates marginal performance, suggesting it cannot discriminate between positive and negative cases (Figures 3,4,5)

Table 1 shows the performance metrics for Protbert, BiLSTM, LSTMAttention, and TextRGNN, with accuracies of 93%, 94%, 93%, and 65%, respectively. The study reveals that ProtBERT and BiLSTM are the best predictors of amino acid sequences for TRPV1-based oral cancer, offering a balance of sensitivity and specificity. TextRGNN, on the other hand, struggles due to low sensitivity and lower accuracy, making it unsuitable for this task. These models could aid in the early detection and treatment of oral cancer. The sensitivity of Protbert, BiLSTM, LSTMAttention, and TextRGNN is the percentage of positive cases correctly detected, while the specificity is the genuine negative rate.

Table 1. Shows the performance metrics for Protbert, BiLSTM, LSTMAttention, and TextRGNN, with accuracies of 93%, 94%, 93%, and 65%, respectively.

Model	ACC	SENSITIVITY	SPECIFICITY	AUC
Protbert	0.9	0.89	0.91	0.937
BiLSTM	0.9	0.91	0.89	0.945
LSTMAttention	0.89	0.89	0.89	0.931
TextRGNN	0.51	0.04	0.98	0.657

The BiLSTM model was the top performer with an AUC of 0.945 and an F1 score of 0.901. Protbert, BiLSTM, and LSTMAttention also showed strong results, with high mean accuracies, sensitivities, specificities, and AUCs of 0.897-0.938. TextRGNN underperformed significantly, achieving only 0.51 in accuracy, 0.04 in sensitivity, 0.98 in specificity, and a poor AUC of 0.657. Strong positive correlations were found between accuracy, sensitivity, AUC, and F1, while specificity negatively correlated, mainly due to TextRGNN's outlier performance.

3. 1 Roc Curve

The Receiver Operating Characteristic (ROC) shows the trade-off between a model's true positive rate (sensitivity) and false positive rate (1-specificity) over categorization thresholds. Protbert, BiLSTM, LSTMAttention, and TextRGNN have high TPR and low FPR in the upper-left ROC region (Figure 3).

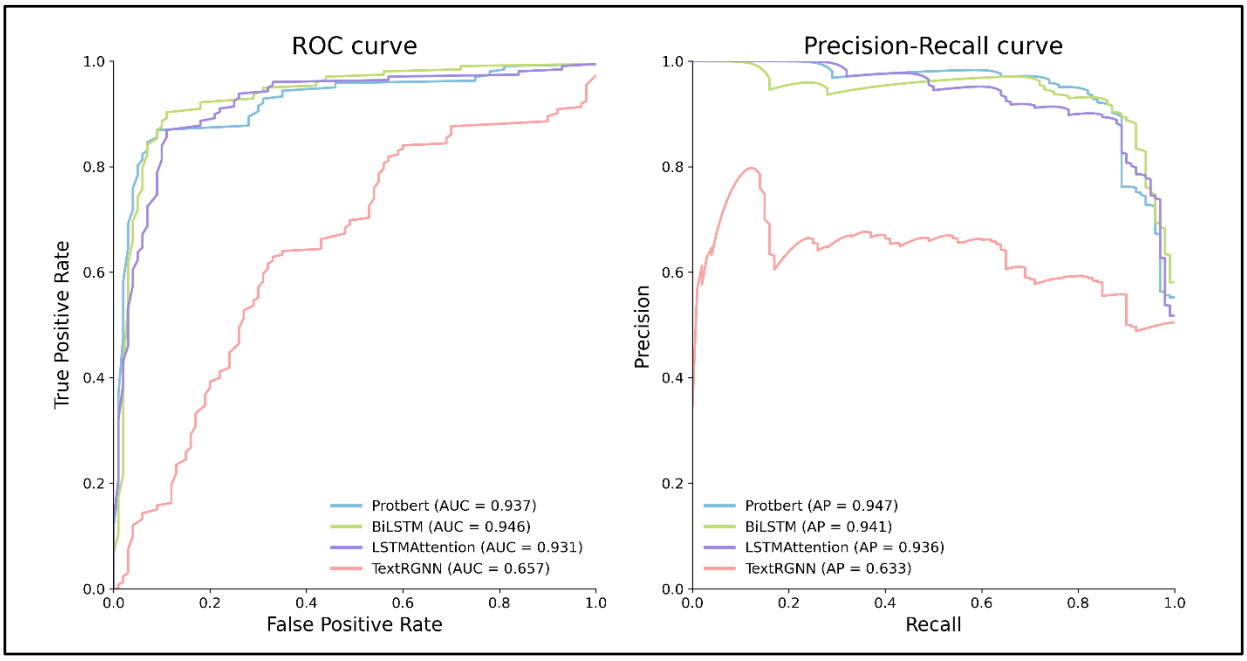

Figure 3. Illustrates the model ROC and precision-recall curves used to evaluate the trade-off between sensitivity and specificity, with a higher AUC indicating better performance. The ROC and Precision–Recall curves compare the classification performance of the four models used in this study. ProtBERT, BiLSTM, and LSTM-Attention consistently exhibit high true-positive detection rates, with AUC values above 0.93, demonstrating their strong ability to differentiate pain-associated TRPV variants. In contrast, TextRGNN performed noticeably worse, as indicated by its lower AUC of 0.657 and decreased precision across recall levels. Overall, the curves show that transformer- and

LSTM-based architectures significantly outperform graph-based models in predicting TRPV variants.

3. 2 Precision Recall Curve

The Receiver Operating Characteristic (ROC) shows the trade-off between a model's true positive rate (sensitivity) and false positive rate (1-specificity) over categorization thresholds. Protbert, BiLSTM, LSTMAttention, and TextRGNN have high TPR and low FPR in the upper-left ROC region (Figure 4).

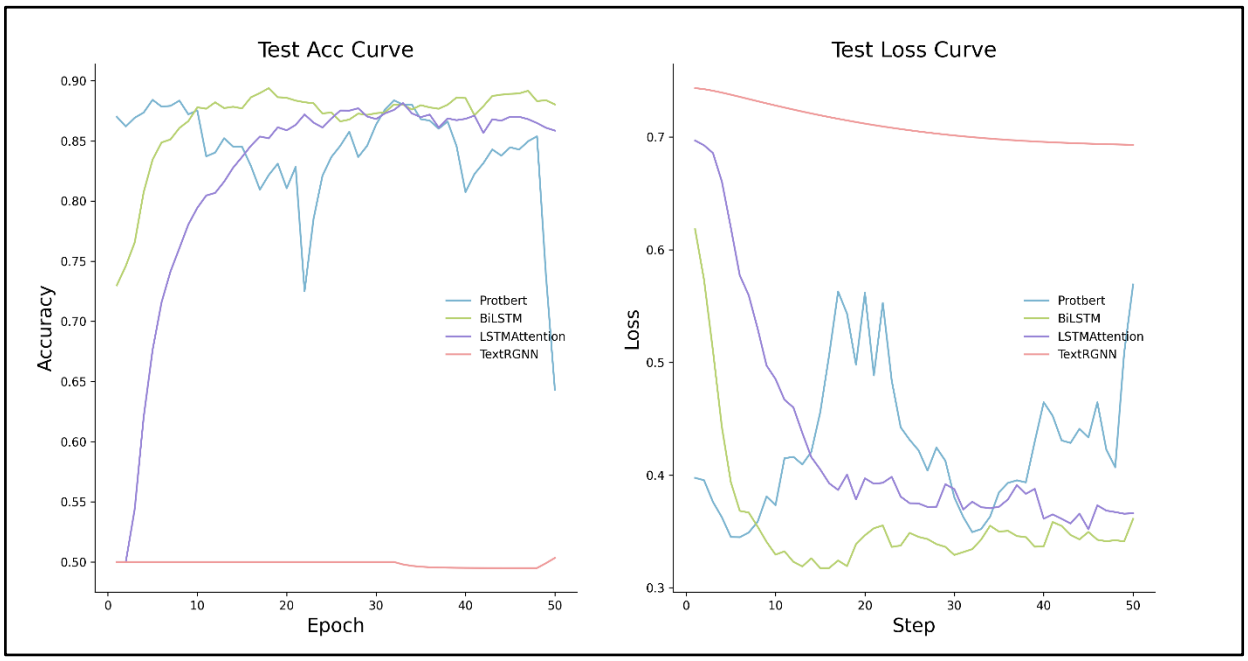

Figure 4. Shows the epoch plots of test accuracy and test loss for the algorithms, and indicates that Protbert, BiLSTM, LSTMAttention, and TextRGNN models achieve moderate accuracy, with peaks around epoch 20. BiLSTM and LSTMAttention show rapid improvement, while TextRGNN remains flat at 0.5, indicating no learning or poor performance. Loss decreases initially but fluctuates after step 20, while TextRGNN shows a slightly higher loss. The test accuracy curves show that ProtBERT, BiLSTM, and LSTM-Attention quickly converge and stay stable around 0.85–0.90, indicating effective learning of TRPV pain-related patterns. In contrast, TextRGNN remains near 0.50, showing it can't learn discriminative features. The test loss curves also decrease and stabilize for the LSTM and transformer models, while TextRGNN's loss stays high. These trends demonstrate that transformer and LSTM models generalize well, but graph-based modeling is insufficient for this dataset.

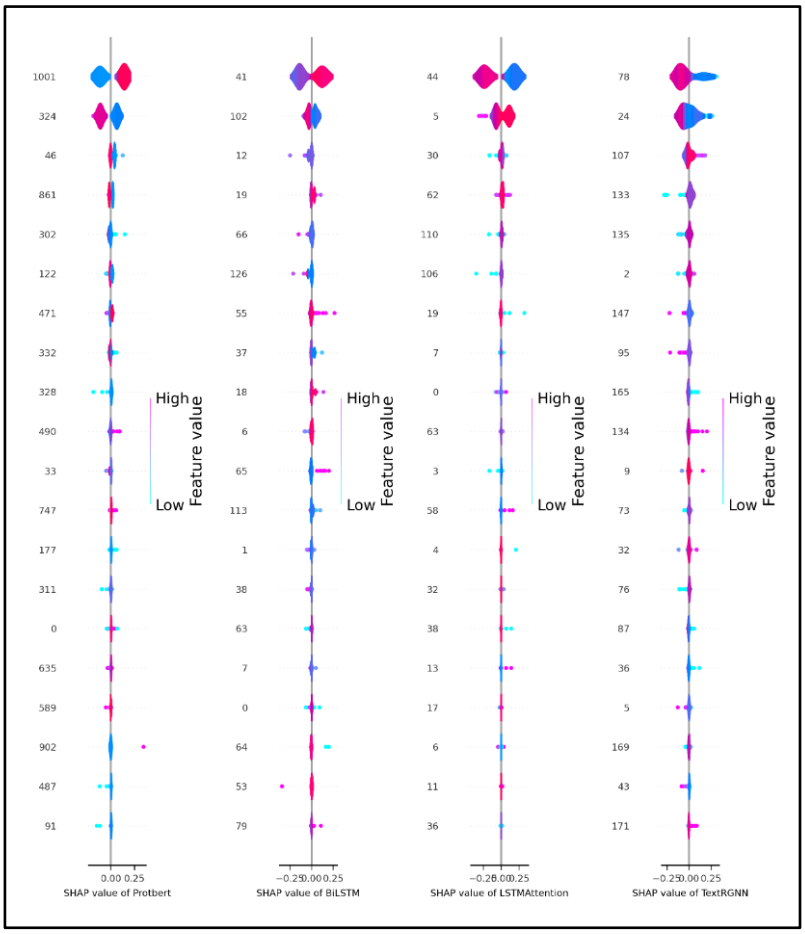

Figure 5. Shows the SHAP values explain model performances and feature contributions. Summary plots show the amino acid features that influence predictions. ProtBERT, BiLSTM, and LSTM-Attention exhibit consistent patterns with high-value features driving positive predictions for pain-related TRPV variants, indicating stable, interpretable learning. TextRGNN produces scattered, low-magnitude SHAP values, indicating weak capture of sequence patterns. Overall, transformer- and LSTM-based models are more accurate and rely on biologically relevant features compared to the graph-based model.

4. Discussion

TRPV1, [6, 7, 32–34] a heat- and ligand-gated ion channel, is pivotal in initiating the pain pathway via sensory ganglion neurons. Transient Receptor Potential Vanilloid (TRPV) proteins constitute a subgroup of the TRP ion channel family. [7, 10, 13, 16, 32] playing diverse roles in sensory perception and cellular processes. In response to heat and capsaicin, TRPV1, TRPV2, TRPV3, and TRPV4 mediate pain and temperature perception. It is implicated in mechanosensation, osmoregulation, and thermosensation. These TRPV proteins are collectively integral to various physiological processes, including nociception, sensory perception, and cellular homeostasis. These protein sequence predictions [27, 33, 34] are essential for designing novel drugs and can reduce resistance. Large protein sequence-based language models can predict protein shapes and activities. Protein design, structure prediction, ligand-binding affinity prediction, and protein-protein interaction prediction are among their molecular biology applications. [23, 30]

The Umami-BERT model, [29] a two-step training method, outperforms other models in predicting umami peptides from amino acid sequences, with high accuracy rates of 93.23% on balanced data and 95.00% on unbalanced data. A recent study using a 738-million-parameter autoregressive Transformer model, ProtGPT2, predicts protein sequences. After training on 50 million non-annotated protein sequences, it can synthesize de novo protein sequences quickly, similar to our study results showing that the BiLSTM, LSTMAttention, and TextRGNN models achieved predictive performance rates of 93%, 94%, and 93%, respectively. The ProtBERT model successfully identifies positive and negative TRPV cases with an accuracy of 0.90, a sensitivity of 0.89, a specificity of 0.91, and an AUC of 0.937. Similarly, the BiLSTM model demonstrates high accuracy and sensitivity, with a specificity of 0.89 and an AUC of 0.945. The LSTM Attention model reliably detects true-positive cases, achieving an accuracy of 0.89, a sensitivity of 0.89, a specificity of 0.89, and an AUC of 0.931 (Figures 3,4,5). Explainable AI models with SHAP plots enhance trust and facilitate informed decision-making in clinical environments. Protbert's color coding distinguishes high and low SHAP values, promoting positive and negative features (Figure 5).

The future of oral cancer pain prediction sequence models includes further fine-tuning existing models and incorporating additional data and features, such as genetic data, clinical parameters, and biochemical markers. Clinical validation studies can assess their practical applicability and foster partnerships between researchers and clinicians. Collaborating with public health organizations can raise awareness about the potential of these predictive models, promote early detection strategies, and empower populations suffering from oral cancer pain. This study improves our understanding of pain mechanisms by accurately predicting functional TRPV sequences, especially TRPV1, paving the way for targeted therapies. TRPV sequences may serve as biomarkers for personalizing pain management strategies, enhancing early detection and diagnosis, streamlining drug discovery and design, and facilitating the integration of AI into clinical practice. The research highlights the critical role of machine learning and AI in clinical workflows for managing pain in cancer patients.

This study has limitations in detecting and treating TRPV1-based oral cancer pain, including data quality, interpretability, overfitting, validation, computational resources, and regulatory issues. Limited datasets may hinder generalization, high sensitivity may cause false positives, and higher costs may result. Overfitting leads to models performing well on training data but poorly on new data, making them unsuitable for clinical use. Computational resources may be scarce in low-resource settings. The study is also limited by annotation-based labels, a lack of wet-lab validation, and potential issues with model generalizability due to sequence diversity. Future work should include larger datasets and experimental validation.

5. Conclusion

Large language models like ProtBERT and BiLSTM could improve TRPV1-based oral cancer pain management, but issues like data quality, interpretability, overfitting, and validation remain. Future research should optimize models, add clinical features, and develop explainable AI tools. These models could help identify pain-related TRPV variants earlier, enabling targeted treatment and better decisions. This study shows ProtBERT and LSTM models can predict pain-related TRPV variants from sequences, indicating potential for biomarker development. However, clinical

use remains speculative without experimental and patient validation.

Abbreviation	Full Form
ProtBERT	A version of BERT (Bidirectional Encoder Representations from Transformers)
BERT	Bidirectional Encoder Representations from Transformers
BiLSTM	Bi-directional Long Short-Term Memory
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
LLM	Large Language Model
SHAP	SHapley Additive exPlanations
TRPV	Transient Receptor Potential Vanilloid

Declarations:

Supplementary Materials: Not applicable.

Author Contributions: Conceptualization, P-K.-Y.-; methodology P-K.-Y.-; software, P-K.-Y.-; validation, P-K.-Y.-; formal analysis, P-K.-Y.-; investigation, P-K.-Y.-; resources, P-K.-Y.-; data curation, P-K.-Y.-; writing- original draft preparation, P-K.-Y.-; writing—review and editing, P-K.-Y.-; visualization, P-K.-Y.-; supervision, P-K.-Y.-; project administration, P-K.-Y.-; funding acquisition. The author has read and agreed to the published version of the manuscript.

Funding: No external funding was received.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Consent for publication: Not applicable.

Data Availability Statement: The data supporting this study's findings are available from the corresponding author upon reasonable request.

Acknowledgments: None.

Conflict of Interests: The author declares no conflict of interest.

References

1. Epstein JB, Miaskowski C. Oral Pain in the Cancer Patient. J Natl Cancer Inst Monogr. 2019;2019(53):lgz003. doi: 10.1093/jncimonographs/lgz003.
2. Matsuka Y. Orofacial Pain: Molecular Mechanisms, Diagnosis, and Treatment 2021. Int J Mol Sci. 2022;23(9):4826. doi: 10.3390/ijms23094826.
3. Y D, Ramani P, Yuwanati M, Ramalingam K, S G. MicroRNA Profiling in Circulating Exosomes in Oral Squamous Cell Carcinoma: A Systematic Review. Cureus. 2023;15(8):e43235. doi: 10.7759/cureus.43235.
4. Hema Shree K, Gayathri R, Veeraraghavan VP, Ramani P, Ramadoss R, Yuwanati M. Gold nanoparticle enhanced TNFα antibody interface using saliva for predicting prognosis in OSCC. Arch Oral Biol. 2025;173:106196. doi: 10.1016/j.archoralbio.2025.106196.
5. Prabhuvenkatesh D, Ramani P, Yuwanati MB, Sukumaran G. Association of Insulin Receptor Substrate 1 Expression in the Prognosis of Head and Neck Squamous Cell Carcinoma: A

Systematic Review and Meta-Analysis. Oncology. 2025;103(3):253-264. doi: 10.1159/000541004.

6. Fernández-Carvajal A, Fernández-Ballester G, Ferrer-Montiel A. TRPV1 in chronic pruritus and pain: Soft modulation as a therapeutic strategy. Front Mol Neurosci. 2022;15:930964. doi: 10.3389/fnmol.2022.930964.
7. Balood M, Ahmadi M, Eichwald T, Ahmadi A, Majdoubi A, Roversi K, et al. Nociceptor neurons affect cancer immunosurveillance. Nature. 2022;611(7935):405-412. doi: 10.1038/s41586-022-05374-w.
8. Dubeykovskaya ZA, Tu NH, Garcia PDR, Schmidt BL, Albertson DG. Oral Cancer Cells Release Vesicles that Cause Pain. Adv Biol (Weinh). 2022;6(9):e2200073. doi: 10.1002/adbi.202200073.
9. Chen L, Wang H, Xing J, Shi X, Huang H, Huang J, Xu C. Silencing P2X7R Alleviates Diabetic Neuropathic Pain Involving TRPV1 via PKCε/P38MAPK/NF-κB Signaling Pathway in Rats. Int J Mol Sci. 2022;23(22):14141. doi: 10.3390/ijms232214141.
10. Muller C, Morales P, Reggio PH. Cannabinoid Ligands Targeting TRP Channels. Front Mol Neurosci. 2019;11:487. doi: 10.3389/fnmol.2018.00487.
11. Løhre ET, Thronæs M, Klepstad P. Breakthrough cancer pain in 2020. Curr Opin Support Palliat Care. 2020;14(2):94-99. doi: 10.1097/SPC.0000000000000494.
12. Bhaskar A. Interventional pain management in patients with cancer-related pain. Postgrad Med. 2020;132(sup3):13-16. doi: 10.1080/00325481.2020.1807796.
13. Khan A, Khan S, Kim YS. Insight into Pain Modulation: Nociceptors Sensitization and Therapeutic Targets. Curr Drug Targets. 2019;20(7):775-788. doi: 10.2174/1389450120666190131114244.
14. Kreutzwiser D, Tawfic QA. Methadone for Pain Management: A Pharmacotherapeutic Review. CNS Drugs. 2020;34(8):827-839. doi: 10.1007/s40263-020-00743-3.
15. Zhang Y, Li X, Liu Y, Li A, Yang X, Tang X. A Multilabel Text Classifier of Cancer Literature at the Publication Level: Methods Study of Medical Text Classification. JMIR Med Inform. 2023;11:e44892. doi: 10.2196/44892. Erratum in: JMIR Med Inform. 2024;12:e62757. doi: 10.2196/62757.
16. Iftinca M, Defaye M, Altier C. TRPV1-Targeted Drugs in Development for Human Pain Conditions. Drugs. 2021;81(1):7-27. doi: 10.1007/s40265-020-01429-2.
17. Luo X, Chen O, Wang Z, Bang S, Ji J, Lee SH, Huh Y, Furutani K, He Q, Tao X, Ko MC, Bortsov A, Donnelly CR, Chen Y, Nackley A, Berta T, Ji RR. IL-23/IL-17A/TRPV1 axis produces mechanical pain via macrophage-sensory neuron crosstalk in female mice. Neuron. 2021;109(17):2691-2706.e5. doi: 10.1016/j.neuron.2021.06.015.
18. Geffen Y, Ofran Y, Unger R. DistilProtBert: a distilled protein language model used to distinguish between real proteins and their randomly shuffled counterparts. Bioinformatics. 2022;38(Suppl_2):ii95-ii98. doi: 10.1093/bioinformatics/btac474.
19. Zhang Y, Zhu G, Li K, Li F, Huang L, Duan M, Zhou F. HLAB: learning the BiLSTM features

from the ProtBert-encoded proteins for the class I HLA-peptide binding prediction. Brief Bioinform. 2022;23(5):bbac173. doi: 10.1093/bib/bbac173.

20. Hallee L, Rafailidis N, Gleghorn JP. cdsBERT - Extending Protein Language Models with Codon Awareness. bioRxiv [Preprint]. 2023:2023.09.15.558027. doi: 10.1101/2023.09.15.558027.
21. Khare E, Gonzalez-Obeso C, Kaplan DL, Buehler MJ. CollagenTransformer: End-to-End Transformer Model to Predict Thermal Stability of Collagen Triple Helices Using an NLP Approach. ACS Biomater Sci Eng. 2022;8(10):4301-4310. doi: 10.1021/acsbiomaterials.2c00737.
22. Wang X, Ding Z, Wang R, Lin X. Deepro-Glu: combination of convolutional neural network and Bi-LSTM models using ProtBert and handcrafted features to identify lysine glutarylation sites. Brief Bioinform. 2023;24(2):bbac631. doi: 10.1093/bib/bbac631.
23. Wang L, Zhong H, Xue Z, Wang Y. Res-Dom: predicting protein domain boundary from sequence using deep residual network and Bi-LSTM. Bioinform Adv. 2022;2(1):vbac060. doi: 10.1093/bioadv/vbac060.
24. Ghazikhani H, Butler G. Enhanced identification of membrane transport proteins: a hybrid approach combining ProtBERT-BFD and convolutional neural networks. J Integr Bioinform. 2023;20(2):20220055. doi: 10.1515/jib-2022-0055.
25. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics. 2022;38(8):2102-2110. doi: 10.1093/bioinformatics/btac020.
26. Wang R, Jiang Y, Jin J, Yin C, Yu H, Wang F, et al. DeepBIO: an automated and interpretable deep-learning platform for high-throughput biological sequence prediction, functional annotation and visualization analysis. Nucleic Acids Res. 2023;51(7):3017-3029. doi: 10.1093/nar/gkad055.
27. Xie YK, Luo H, Zhang SX, Chen XY, Guo R, Qiu XY, et al. GPR177 in A-fiber sensory neurons drives diabetic neuropathic pain via WNT-mediated TRPV1 activation. Sci Transl Med. 2022;14(639):eabh2557. doi: 10.1126/scitranslmed.abh2557.
28. Zhang W, Lyu M, Bessman NJ, Xie Z, Arifuzzaman M, Yano H, Parkhurst CN, Chu C, Zhou L, Putzel GG, Li TT, Jin WB, Zhou J; JRI Live Cell Bank; Hu H, Tsou AM, Guo CJ, Artis D. Gut-innervating nociceptors regulate the intestinal microbiota to promote tissue protection. Cell. 2022;185(22):4170-4189.e20. doi: 10.1016/j.cell.2022.09.008.
29. Xie Z, Feng J, Hibberd TJ, Chen BN, Zhao Y, Zang K, Hu X, Yang X, Chen L, Brookes SJ, Spencer NJ, Hu H. Piezo2 channels expressed by colon-innervating TRPV1-lineage neurons mediate visceral mechanical hypersensitivity. Neuron. 2023;111(4):526-538.e4. doi: 10.1016/j.neuron.2022.11.015.
30. Mailhot B, Christin M, Tessandier N, Sotoudeh C, Bretheau F, Turmel R, et al. Neuronal interleukin-1 receptors mediate pain in chronic inflammatory diseases. J Exp Med. 2020;217(9):e20191430. doi: 10.1084/jem.20191430.
31. Littmann M, Bordin N, Heinzinger M, Schütze K, Dallago C, Orengo C, Rost B. Clustering

FunFams using sequence embeddings improves EC purity. Bioinformatics. 2021;37(20):3449-3455. doi: 10.1093/bioinformatics/btab371.

32. Hu J, Zeng WW, Jia NX, Arif M, Yu DJ, Zhang GJ. Improving DNA-Binding Protein Prediction Using Three-Part Sequence-Order Feature Extraction and a Deep Neural Network Algorithm. J Chem Inf Model. 2023;63(3):1044-1057. doi: 10.1021/acs.jcim.2c00943.
33. Xu T, Wang Q, Yang Z, Ying J. A BERT-based approach for identifying anti-inflammatory peptides using sequence information. Heliyon. 2024;10(12):e32951. doi: 10.1016/j.heliyon.2024.e32951.
34. Jha K, Karmakar S, Saha S. Graph-BERT and language model-based framework for protein-protein interaction identification. Sci Rep. 2023;13(1):5663. doi: 10.1038/s41598-023-31612-w.
35. Zhang J, Yan W, Zhang Q, Li Z, Liang L, Zuo M, Zhang Y. Umami-BERT: An interpretable BERT-based model for umami peptides prediction. Food Res Int. 2023;172:113142. doi: 10.1016/j.foodres.2023.113142.