學 士 學 位 論 文

난소암 진단 모델 구축을 위한 자연어 처리 기반 RNA-seq 데이터 전처리 방법 연구

Language-Model-Based RNA-seq Data ETL Workflow for Ovarian Cancer Classification Model

2025年 6 月

韓 東 大 學 校

生命科學部

구나영 (具 奈 映)

난소암 진단 모델 구축을 위한 자연어 처리 기반 RNA-seq 데이터 전처리 방법 연구

지도교수 안 태 진

이 논문을 학사학위 논문으로 제출합니다.

2025年 6月

한 동 대 학 교

생명과학부

22101006

구 나 영 (具 奈 映)

구 나 영의 학사학위 논문을 인준합니다.

2025년 6월

지도교수 안 태 진 (서명)
학 부 장 백 재 현 (서명)

목 차

Abstract

Early and accurate diagnosis of ovarian cancer remains a major clinical challenge due to the lack of reliable biomarkers and the complexity of transcriptomic data. While RNA-seq provides rich molecular information, conventional analysis pipelines often cannot fully leverage the sequential and contextual nature of biological sequences.

In this study, I propose a novel ETL pipeline that treats RNA-seq raw data as a biological language. Raw nucleotide sequences are tokenized at the gene level and translated into amino acid sequences, embedded using a pretrained transformer model (ProtBERT), and classified using a Bi-LSTM network to distinguish cancerous from non-cancerous samples.

My approach achieves high diagnostic performance, with an AUC of 0.7067 in the test set, demonstrating the potential of contextual embeddings to capture disease-related signals. Beyond classification, the generated embeddings can provide meaningful insight to utilize a language model and natural language processing in sequencing data analysis.

I. Introduction

Ovarian cancer remains one of the deadliest gynecologic malignancies due to late detection. Because of vague symptoms in the early stage, an effective diagnostic method is essential. Serum cancer antigen 125 (CA125) and human epididymis protein 4 (HE4) are important biomarkers used to monitor and screen ovarian cancer; however, they show conflicting results with a high false-positive rate [1, 2, 3]. As extracellular RNA in blood and biofluids has emerged as a noninvasive biomarker for various diseases, RNA sequencing (RNA-Seq) has become a powerful tool for profiling gene expression patterns in specific diseases [4].

However, although RNA-seq analysis provides underlying pathogenesis and heterogeneity, the traditional preprocessing pipeline for RNA-seq data typically relies on statistical and quantitative approaches that may fail to capture the contextual information in genomic sequences [5]. Transcriptomes from blood-based liquid biopsies are particularly high-dimensional and sparse, so gene expression quantification methods have yet to reach a consensus. Methods that can contain contextual dependency remain underexplored.

Genomic sequences, like natural language, encode meaning and function through their specific sequences [6]. In particular, nucleotides encode amino acid sequences through quaternary structures, and the amino acid sequence encodes the protein's function and properties through vigesimal codes. As this analogy shows, the possibility of applying natural language processing (NLP) in omics data, I hypothesized that treating protein sequence as a biological language could contain compressed patterns that can distinguish between healthy and cancerous samples.

Among various omics data, platelet-derived RNA-seq is promising data for cancer diagnosis, as platelets circulate throughout the body and their RNA profile contains a tumor-associated RNA pattern [7]. Motivated by the structural similarity between genomic sequences and natural language, I propose a novel approach that treats RNA-seq data as a biological corpus with context, applying a large language model-based embedding to capture the semantic features of amino acid sequences translated from RNA [8].

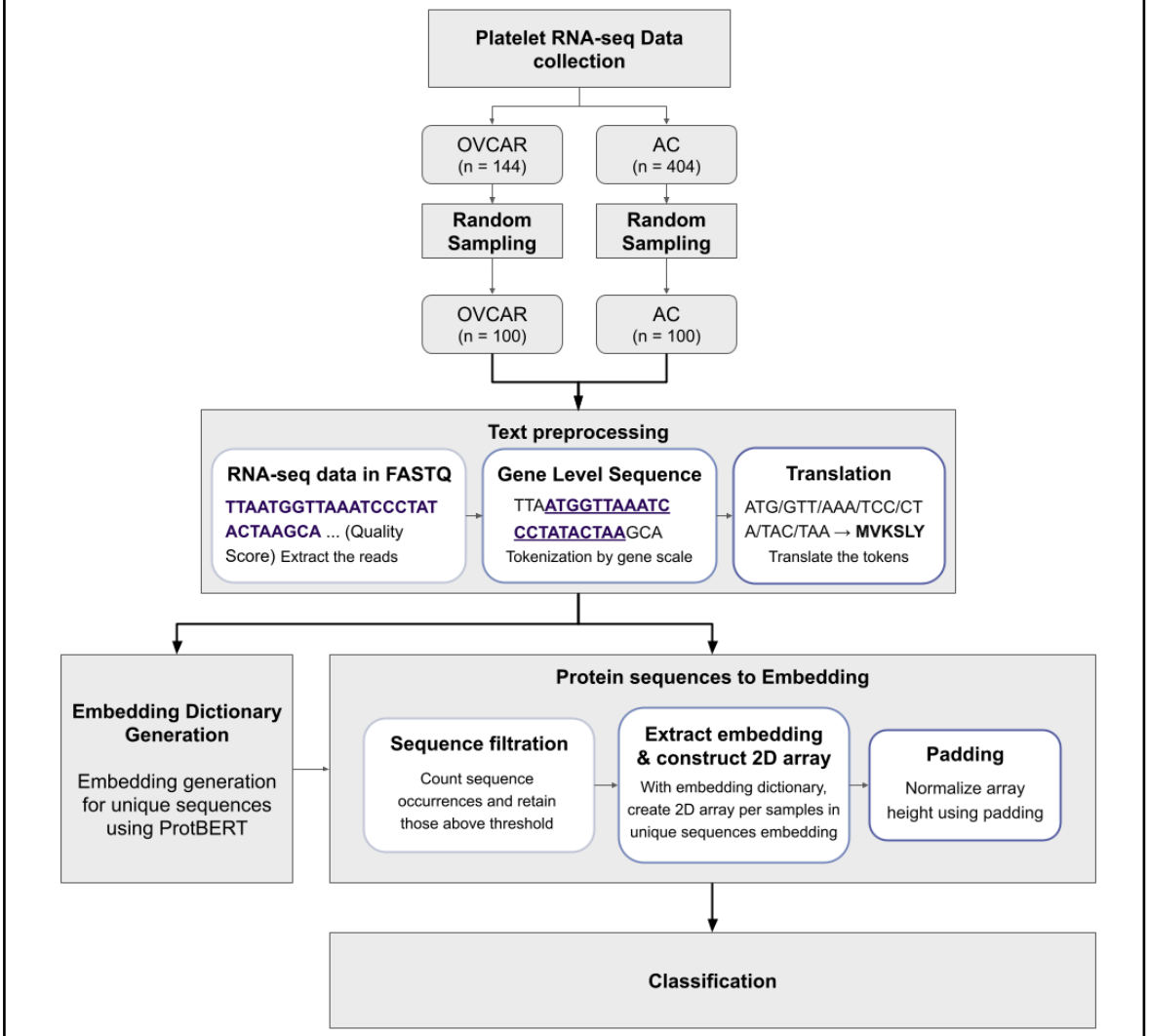
In this study, I present a novel ETL pipeline to preprocess RNA-seq raw data within an NLP framework. Specifically, I convert nucleotide sequences derived from platelet RNA-seq into amino acid sequences, then apply a pretrained language model, ProtBERT, to generate contextual embedding for each sequence to make sure the embeddings can serve as high-dimensional representations that capture the semantic and structural information of protein sequences. I utilized the embeddings as input data to a deep learning-based diagnostic model to classify cancerous and

non-cancerous samples, aiming to extract patterns and insights that could reveal new ways to diagnose diseases. This approach demonstrates the potential of applying NLP-based embeddings to biological sequence data such as RNA-seq.

II. Results

1. Workflow of data preparation

Figure 1. Overview of the RNA-seq to Protein Embedding and Classification Pipeline



To build a cancer classification model based on platelet RNA-seq data, a multi-step preprocessing pipeline was designed, similar to natural language processing that converts raw 1D sequencing reads into standardized 1028-D protein embedding arrays:

(1) Sample Preparation: Platelet RNA-seq data were collected from two groups — OVCAR and controls. I randomly sampled 100 individuals from each group to balance the dataset.
(2) Text-based Preprocessing: Raw FASTQ reads were processed to extract nucleotide

sequences. These were segmented into gene-level tokens and translated into amino acid sequences.

(3) Embedding Generation: All unique amino acid sequences were passed through ProtBERT to generate high-dimensional embeddings. A dictionary mapping each sequence to its embedding vector was constructed. (More details on the Methods)
(4) Sequence Filtering: For each sample, the occurrence of protein sequences was counted and retained with frequencies above a threshold (cutline = 50 or 100) to emphasize biologically relevant patterns.
(5) Embedding Matrix Construction: Using the embedding dictionary, I constructed a 2D array of shape (number of unique sequences, 1024) for each sample, containing only the selected high-frequency protein embeddings.
(6) Normalization via Padding: To ensure uniform input size for modeling, each array was vertically (by the number of unique sequences) padded to the maximum, mean, or median observed sequence count of all samples.

Multiple datasets were generated by varying preprocessing parameters (Table 1). The final dataset, consisting of shape (200 for the sum of sample number, number of unique sequences, 1024 for embedding dimension) with corresponding binary labels (normal vs. cancer), was used as input for downstream deep learning classification. The dataset in Table 1 was selected because it yielded the highest classification performance among them. (More details on the workflow are described in Figure 1 and the Methods.) In particular for embedding generation, once amino acid sequences are made, ProtBERT was used—a pre-trained language model for proteins—to generate 1024-dimensional embeddings [9]. These embeddings capture the context and properties of each sequence. To streamline the process, a dictionary mapping unique sequences to their embeddings was created, allowing efficient retrieval during analysis.

Table 1. The generated dataset. 200 is for the dataset sample size, and 1028 is the embedding dimension

cutline	padding criterion	shape (padded) of each sample	dataset name
50	max	200, 2461, 1028	data/v2_50_max_2461.npz
100	max	200, 1261, 1028	data/v2_100_max_1261.npz
50	median	200, 50, 1028	data/v2_50_max_2461.npz
100	median	200, 292, 1028	data/v2_50_max_2461.npz
50	mean	200, 686, 1028	data/v2_50_mean_686.npz
100	mean	200, 340, 1028	data/v2_100_mean_340.npz

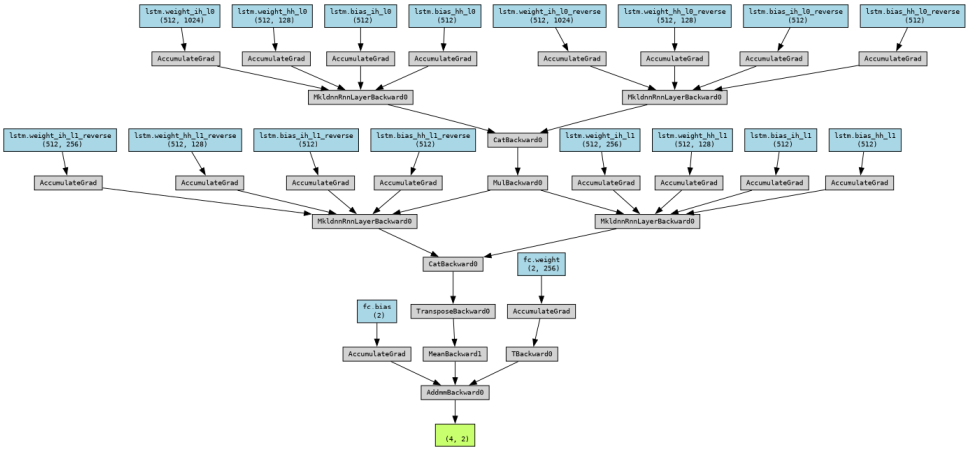
2. Best Model for Classification using LSTM

Figure 2B is the AUC-ROC curve for our best-performing model, which is an LSTM-architected neural network. The best performance with this architecture was achieved by the dataset whose cutline for sequence filtering was 50, and the embedding array was padded by the maximum number, so that its shape per sample is 2461 and 1027 for the number of unique sequences and embeddings. The LSTM model emerged as the best performer, achieving an AUC of 0.7067 in the test set.

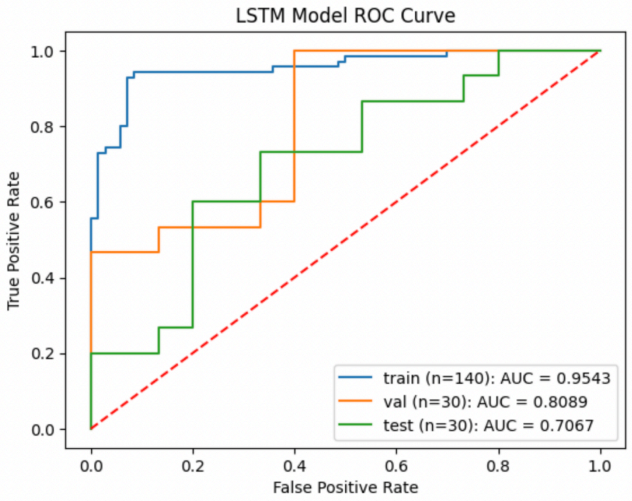
Figures 2C, 2D, and 2Eshow the confusion matrices for the model's result. While the train set and the test show the same size of false positive (FP) and false negative (FN), the validation data shows one more FP case than the FN cases are diagnosed as cancer.

Figure 2. Best Model

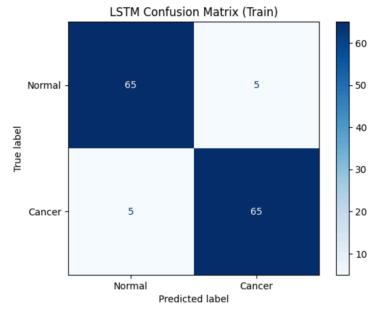
(A)



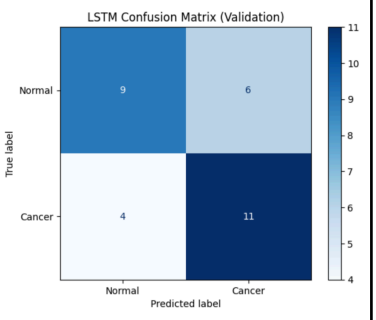
(B)



(C)



(D)



(E)

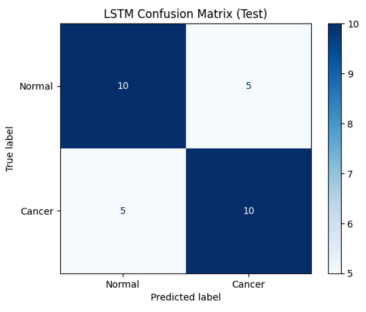


Figure 2. Best Model

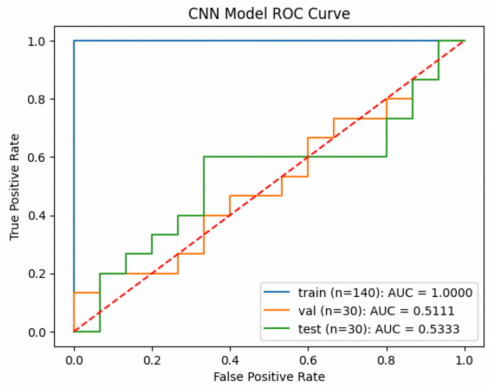
Figure 2 shows the result of LSTM with data/v2_50_max_2461.npz. Train AUC = 0.9543; Val AUC = 0.8089; Test AUC = 0.7067 (A) bi-LSTM architecture (B) AUC ROC Curve (C) Confusion matrix of train set (D) Confusion matrix of validation set (E) Confusion matrix of test set.

3. Other Classification Model Results using Different Neural Network Architectures

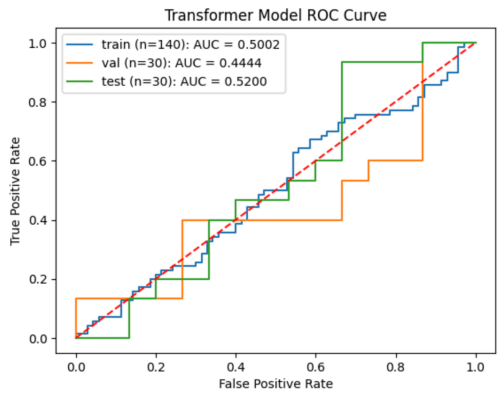
For the other classification architecture, CNN, Transformer, and Attention were deployed. And I trained them using different datasets. Figure 3D is the AUC-ROC curve for the best-performing model using LSTM. The rest of Figure 3 shows the other architectures' best score by the test data AUC. The dataset for each best score is in Table 2. While the LSTM model achieved an AUC of 0.7 in the test set, the CNN, Transformer, and Attention models show lower test AUCs, ranging from 0.3333 to 0.64 (Figure 3A, 3B, 3C and Table 2).

Figure 3. Other Classification Architectures' Best Score (by test data AUC)

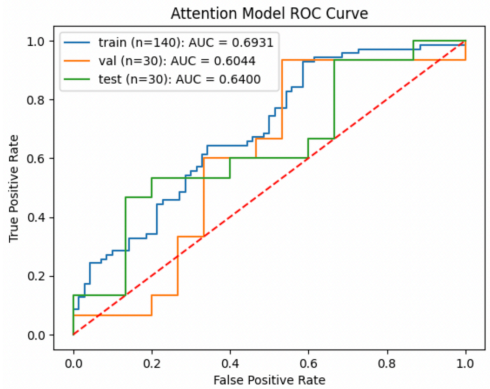
(A)



(B)



(C)



(D)

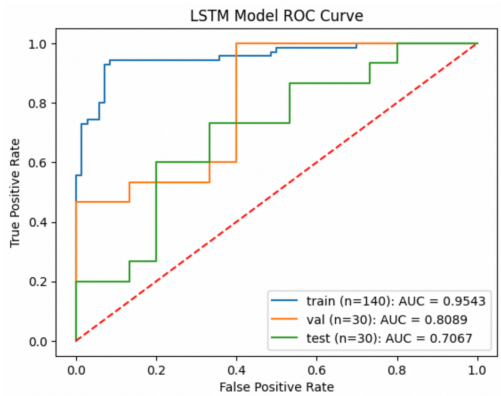


Table 2. Other Classification Architectures' Best Score (by test data AUC)					
Figure		3A	3B	3C	3D
Architecture		CNN	Transformer	Attention	LSTM
name of dataset		data/v2_100_max_1261.npz	data/v2_50_median_592.npz	data/v2_100_max_1261.npz	data/v2_50_max_2461.npz
AUC	train	1.0	0.5002	0.6931	0.9543
	val	0.5111	0.4444	0.6044	0.8089
	test	0.5333	0.52	0.64	0.7067

4. Other LSTM-based Classification Model Results with Different Datasets

To check whether the model is affected by input data shape, I utilized every dataset in different permutations of parameters and checked the performance by training them in LSTM, the best model. As Figure 4 and Table 3 show how the model performance varies by the input. Figure 4F is the best-performing input model, which was also in Figures 2B and 3D.

Figure 4. LSTM with other datasets

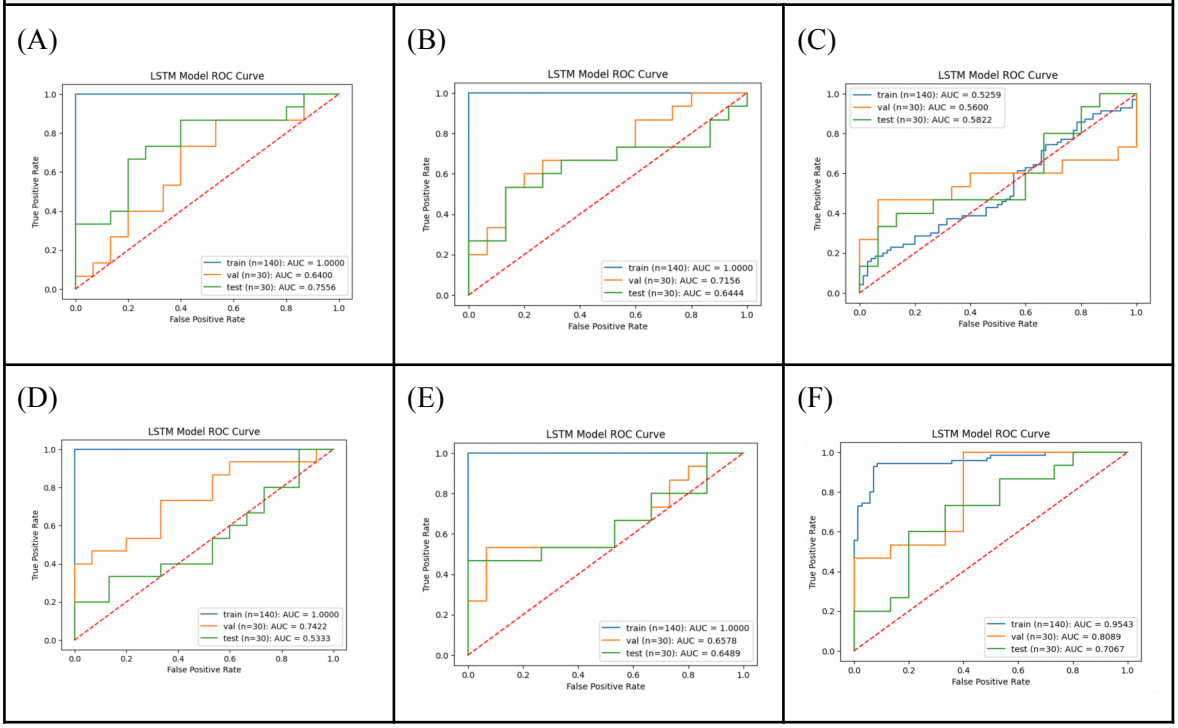


Table 3. LSTM with other datasets						
Figure	4A	4B	4C	4D	4E	4F

Table 3. LSTM with other datasets							
cutline		100			50		
padded by		median	mean	max	median	mean	max
n		292	340	1261	592	686	2461
AUC	train	1.0	1.0	0.5259	1.0	1.0	0.9543
	val	0.64	0.7156	0.56	0.7422	0.6578	0.8089
	test	0.7556	0.6444	0.5822	0.5333	0.6489	0.7067

III. Discussion

Biological sequences have been preprocessed by the protocol based on a quantitative approach. Here, I suggest natural language processing (NLP) based preprocessing in RNA-seq as a hybrid of the qualitative analysis power of NLP and quantitative ability of RNA-seq data by showing that it can detect the difference between ovarian cancer and non-cancer patients. LSTM's ability to generalize better than other architectures underscores its suitability for genomic sequence analysis. The models demonstrate the potential of applying deep learning to genomic data. The best model in Figure 2 shows an AUC of 0.7067 in the test set, which demonstrates its potential for distinguishing between cancerous and normal samples. This model's neural network architecture was based on LSTM, and the input dataset's length was the largest. This result suggests that LSTM's ability to capture sequential dependencies is well-suited for genomic data. Figure 3-A, B, C, and Table 2 showing the other models with the lower AUC values than LSTM also highlight that sequential order in genomic data plays a critical role.

While the model successfully identified most cancerous and normal samples, some misclassifications occurred in the validation set. The drop from training (AUC=0.9543) and validation (AUC=0.8089) to testing indicates potential overfitting or a need for more diverse test data. Many of these errors were associated with noisy or ambiguous sequences, which could be addressed in future improvements.

Also, the limited dataset size may have restricted the ability to achieve a higher test AUC. CNN, Transformer, and Attention models require further optimization to match the performance of LSTM. In addition, Figure 4 shows variability in AUC scores across datasets; Performance is sensitive to input characteristics, such as embedding cutlines. Thus, it might be struggling to find the optimal cutline and the padding criteria, considering generalizability.

In conclusion, this project explored how principles from natural language processing can revolutionize genomic data analysis. By leveraging ProtBERT embeddings and hybrid models, the workflow achieved meaningful insights into cancer diagnosis. While this study marks an early step, it demonstrates the feasibility and promise of combining AI with genomics for both research and clinical applications.

IV. Methods

1. Data Preparation

RNA-seq datasets were collected from two cohorts: ovarian cancer patients and healthy controls. I started with RNA-seq data in FASTQ format from ovarian cancer and normal samples. Blood platelets RNA-seq data. The data was given by the Biodata Lab at Handong Global

University.

To make this data usable for our approach, I first split it into gene-level sequences by identifying start and stop codons. Next, I translated these gene sequences into amino acid sequences, treating them as functional units analogous to sentences in natural language.

2. Embedding Dictionary Generation

Once I had the amino acid sequences, I used ProtBERT—a pre-trained language model for proteins—to generate 1024-dimensional embeddings. These embeddings capture the context and properties of each sequence. To streamline the process, I created a dictionary mapping unique sequences to their embeddings, allowing efficient retrieval during analysis.

ProtBERT was trained on datasets like UniRef100 and BFD, which contain a vast array of protein sequences [9]. This diverse training data contributes to the model's robustness and ability to generalize across different protein-related tasks. The model utilizes self-supervised learning techniques, allowing it to learn from large amounts of unlabeled protein data. ProtBERT employs a masked language modeling approach similar to BERT. The model is trained to predict these masked tokens based on the context provided by the surrounding unmasked tokens. This approach helps ProtBERT to understand the underlying structure and relationships within protein sequences, indicating that this model understands the "grammar" of protein sequences, enhancing its predictive capabilities for various biological tasks.

So I used ProtBERT to generate 1024-dimensional embeddings with the amino acids. These embeddings capture the context and properties of each sequence. To streamline the process, I created a dictionary mapping unique sequences to their embeddings, allowing efficient retrieval during analysis. So I first randomly sampled 10 from each group (cancer and non-cancer) and extracted unique amino acid sequences whose length is more than 20. Then I made two embedding dictionaries for each group: the key was the amino acid sequence, and the values were the embedding vectors.

3. Amino Acid Sequences to Embedding

Using the dictionary for each group, the sample data in each group was also converted into an embedding. Then the datasets were padded to ensure uniform input size across samples. This step allowed us to combine quantitative counts with qualitative sequence patterns in our analysis. With this process, I made 200 samples with a 2D array, but since its number of rows is different, the dataset should be padded. So I made 6 datasets with different permutations in parameters: cutline of repeat filtering as 50 and 100, which are two possibilities; and padding criteria, mean, median, or max (Table 1).

4. Training and Validation

For training, I split the data (n=200) into training, validation, and test sets, using a stratified approach to balance the labels. 70% was used as a training set (n=140), 15% each for validation and test sets (n=30, 30). All the features were scaled by the Standard Scaler. The dataset in the numpy format was changed into a Tensor. Models were trained using the Adam (Adaptive Moment Estimation) optimizer with a learning rate scheduler called ReduceLROnPlateau for 10 epochs and a batch size of 4.

5. Classification Models

These embedding arrays for each sample were tested by four different neural network architectures: Convolutional Neural Network(CNN), bidirectional Long Short-Term Memory

(bi-LSTM), Transformer, and Attention. I experimented with CNNs, LSTMs, and Transformer-based architectures, each leveraging the contextual embeddings to learn patterns indicative of cancer. The model outputs were binary classifications—cancerous or normal—evaluated using the AUC-ROC metric to measure performance.

Table 4. Neural Network Characteristics

Models	**Characteristics**	**# of parameters**
bidirectional LSTM (bi-LSTM)	a model that handles sequential dependencies by learning long-term and short-term patterns	1,577,474
Transformer [10]	a model that processes sequences using self-attention, allowing the model to weigh the importance of each part of the sequence regardless of its distance from other parts	18,279,938
Multi-head Attention	a model that uses multiple attention mechanisms to focus on different parts simultaneously, extracting diverse contextual information, also part of a transformer component	4,331,906
CNN	A model that applies convolutional filters to extract local patterns in data, often used for image and pattern analysis.	10,273,218

6. Computing Resources

All the data preprocessing and dictionary generation processes were executed in Biodata Laboratory's SSH server, especially bds3. All the classification processes were executed in HGUCOSS from the Big Data Hub at Handong Global University.

IV. References

1. Fawzy, A., Mohamed, M.R., Ali, M.A., El-Magied, M.H.A. and Helal, A.M. (2016). Tissue CA125 and HE4 Gene Expression Levels Offer Superior Accuracy in Discriminating Benign from Malignant Pelvic Masses. *Asian Pacific Journal of Cancer Prevention*, 17(1), pp.323–333.
2. Akinwunmi, B.O., Babic, A., Vitonis, A.F., Cramer, D.W., Titus, L., Tworoger, S.S. and Terry, K.L. (2018). Chronic Medical Conditions and CA125 Levels among Women without Ovarian Cancer. *Cancer Epidemiology Biomarkers & Prevention*, 27(12), pp.1483–1490.
3. Zhang, M., Cheng, S., Jin, Y., Zhao, Y. and Wang, Y. (2021). Roles of CA125 in diagnosis, prediction, and oncogenesis of ovarian cancer. *Biochimica et Biophysica Acta (BBA) - Reviews on Cancer*, 1875(2), p.188503.
4. Hulstaert, E., Morlion, A., Levanon, K., Vandesompele, J. and Mestdagh, P. (2021). Candidate RNA biomarkers in biofluids for early diagnosis of ovarian cancer: A systematic review. *Gynecologic Oncology*, [online] 160(2), pp.633–642.
5. Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., Szcześniak, M.W., Gaffney, D.J., Elo, L.L., Zhang, X. and Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. *Genome Biology*, 17(1).
6. Rannon, E. and Burstein, D. (2025). *Leveraging Natural Language Processing to Unravel the Mystery of Life: A Review of NLP Approaches in Genomics, Transcriptomics, and Proteomics.*

[online] arXiv.org. Available at: https://arxiv.org/abs/2506.02212.
7. Best, Myron G., Sol, N., Kooi, I., Tannous, J., Westerman, Bart A., Rustenburg, F., Schellen, P., Verschueren, H., Post, E., Koster, J., Ylstra, B., Ameziane, N., Dorsman, J., Smit, Egbert F., Verheul, Henk M., Noske, David P., Reijneveld, Jaap C., Nilsson, R. Jonas A., Tannous, Bakhos A. and Wesseling, P. (2015). RNA-Seq of Tumor-Educated Platelets Enables Blood-Based Pan-Cancer, Multiclass, and Molecular Pathway Cancer Diagnostics. *Cancer Cell*, 28(5), pp.666–676.
8. Ji, Y., Zhou, Z., Liu, H. and Davuluri, R.V. (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15).
9. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D. and Rost, B. (2021). ProtTrans: Towards Cracking the Language of Lifes Code Through Self-Supervised Deep Learning and High Performance Computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10), pp.1–1.
10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I. (2017). *Attention Is All You Need*. [online] arXiv. Available at: https://arxiv.org/abs/1706.03762.