Biomedical Text Classification using LSTM, GRU and Bahdanau Attention

Mvomo Eto Wilfried
Master in Data Science and Engineering
University of Liège

Promotor : Ashwin Ittoo

Academic Year : 2024-2025

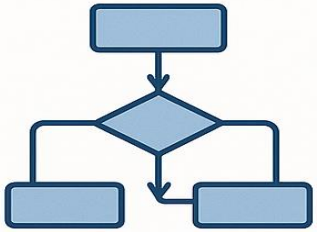
Outline

1. Introduction & Context

2. Research Questions

3. Data & Preprocessing

4. Methodology



5. Experiments & Key Results

6. Conclusion & Perspectives

1.Introduction & Context

- Biomedical NLP is crucial for clinical decision support, literature mining and disease surveillance;
- Challenges include specialized vocabulary and limited annotated data;

2.Research Questions

Q.1 How does training time vary across different model architectures and embedding strategies, and what trade-offs exist between performance and computational efficiency?

Q.2 What is the impact of different class imbalance mitigation strategies, such as SMOTE versus class weighting, on model performance across resource disease categories?

Q.3 And to what extent do the best-performing models generalize to underrepresented disease categories, and can few-shot learning methods effectively improve classification in low-resource scenarios?

3.Data & Preprocessing

Datasets

- 1. **Memorial Sloan Kettering Cancer Center Dataset** available at https://www.kaggle.com/competitions/msk-redefining-cancer-treatment (~ 3 500 abstracts);
- 2. **Medical Abstracts Dataset** available at https://github.com/sebischair/Medical-Abstracts-TC-Corpus (~ 12 000 abstracts);

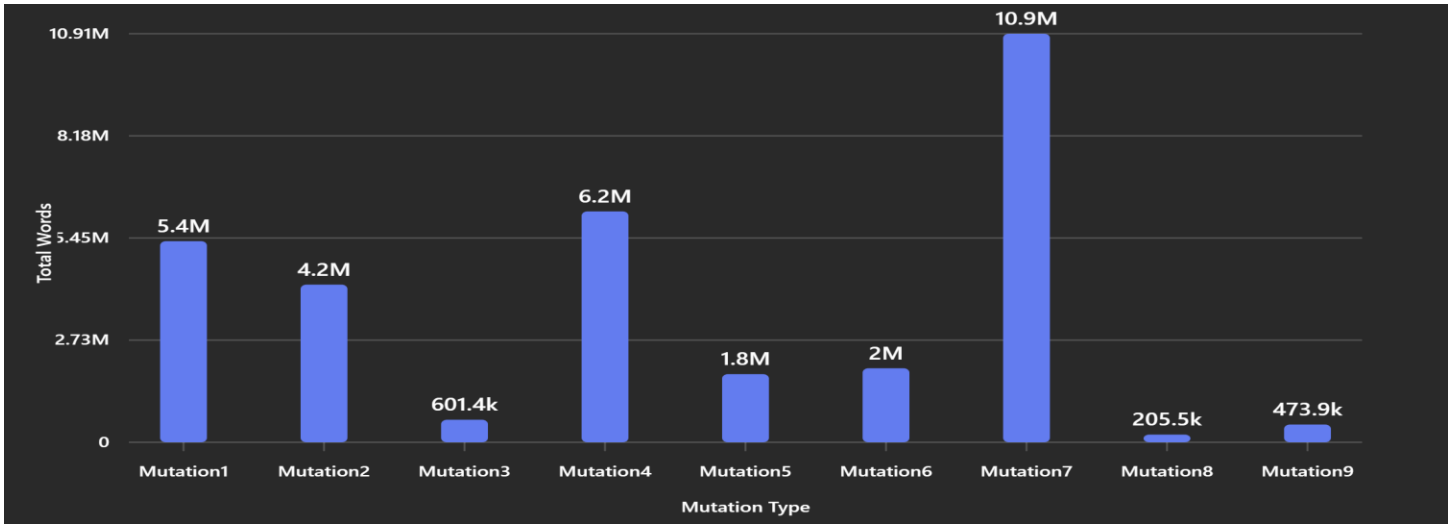
Structure of the MSK Kaggle Challenge Dataset

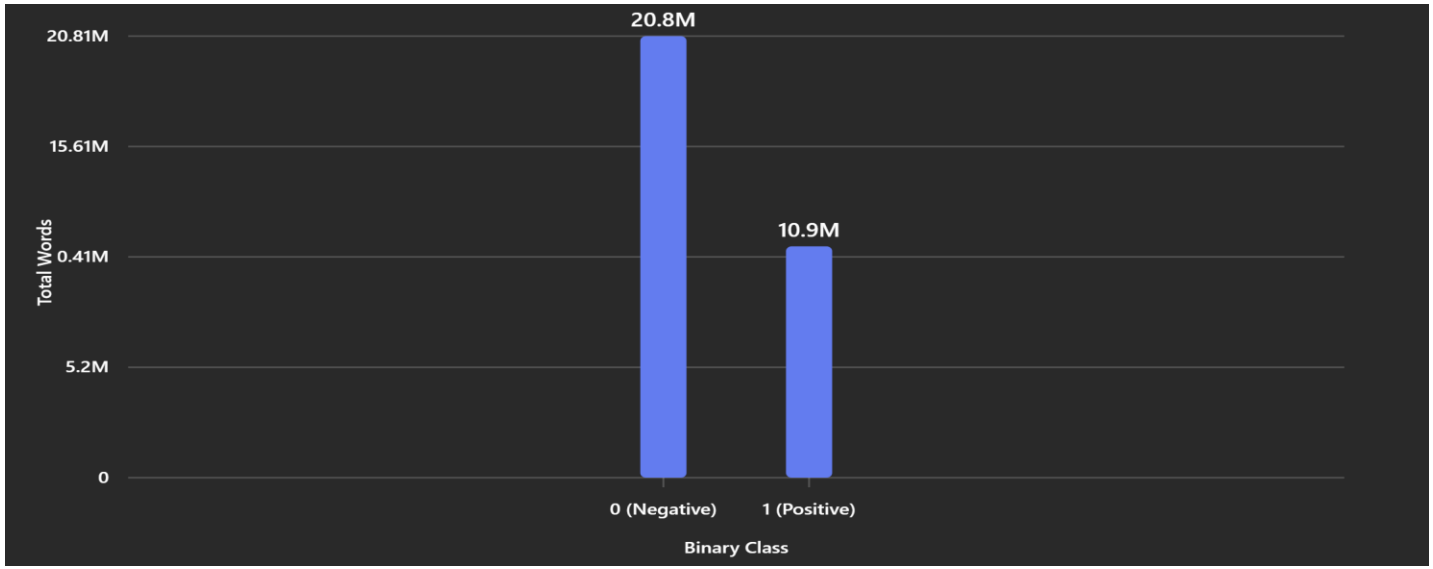
Variable	Unique identifier for linking with clinical evidence (in training_variants) or matching the mutation metadata (in training_text).
Gene	Gene in which the mutation occurs.
Variation	Amino acid change caused by the mutation.
Class	Integer (1–9) indicating the clinical relevance category.
Text	Clinical description and literature excerpts used for classification

Structure of the Medical Abstracts Dataset

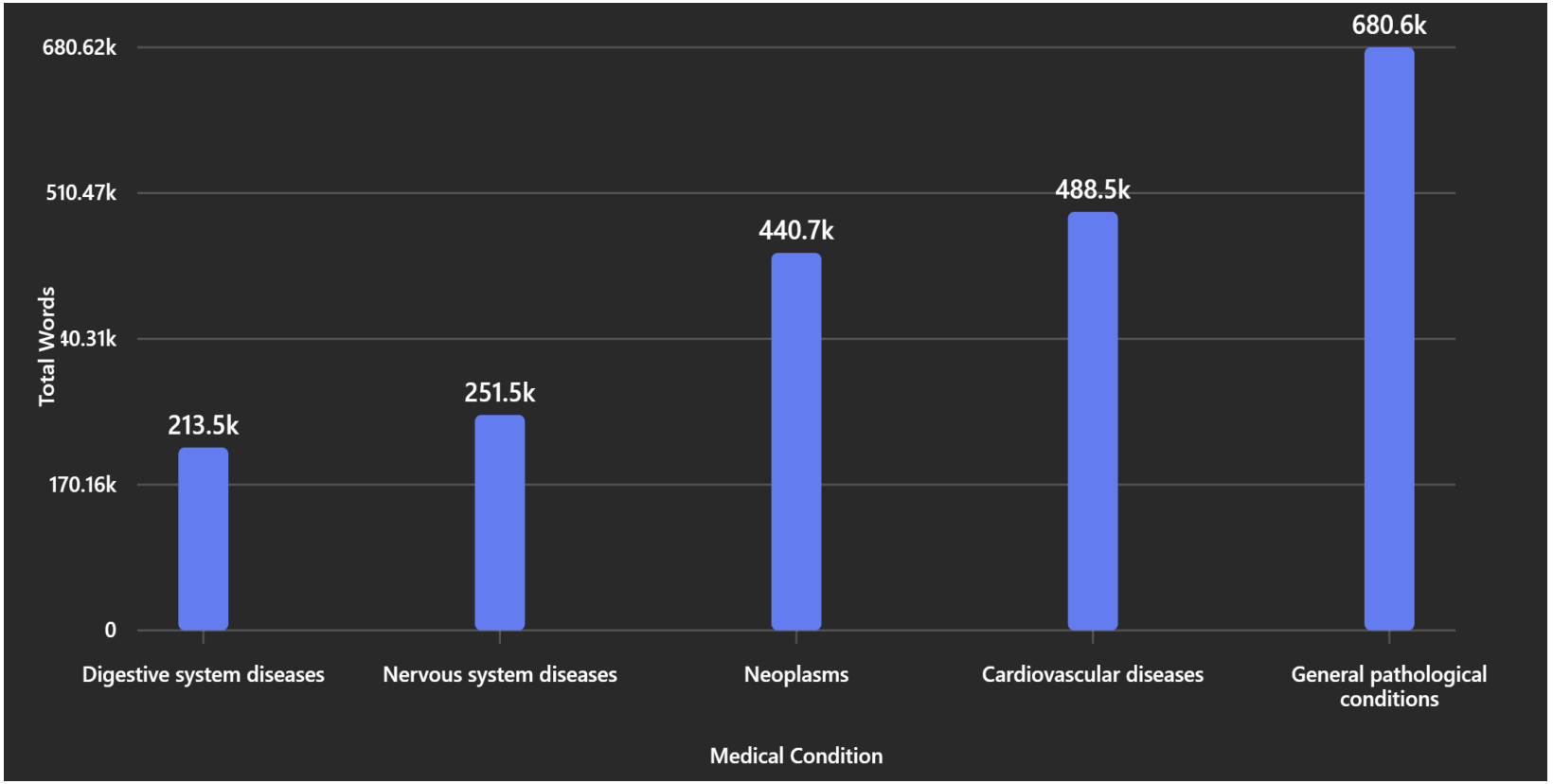
Variable	Description
condition_ label	Integer (1–5) representing the patient condition category: 1 = Neoplasms, 2 = Digestive systm diseases, 3 = Nervous system diseases, 4 = Cardiovascular diseases, 5 = General pathological conditions.
medical_ abstract	Text of the medical abstract summarizing clinical information related to the patient condition.
Variable	**Description**
condition label	Teteger (1–5) representing the patient condition category

Memorial Sloan Kettering Cancer Center Dataset: Mutation 7 (positive class) vs Rest (negative class).





Medical Abstracts Dataset : 0: Neoplasms, **1:** Digestive system diseases, **2:** Nervous system diseases, **3:** Cardiovascular diseases and **4:** General pathological conditions.



4.Methodology

Machine Learning Models :

- Linear Support Vector Machine ;
- Random Forest;

Deep Learning Models :

- Gated Recurrent Unit (GRU);
- Long Short-Term Memory (LSTM);
- Incorporating Bahdanau attention;

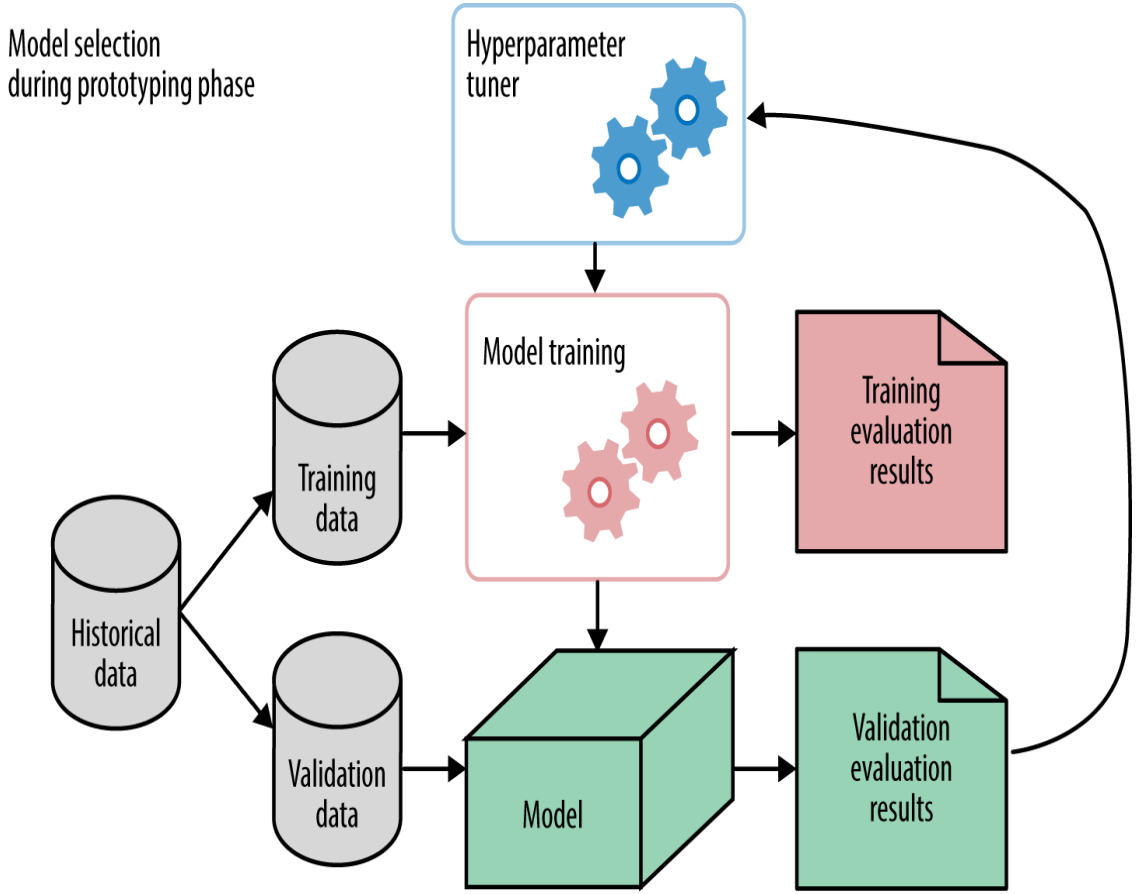
Embedding Techniques :

- Learned from scratch;
- Word Embeddings : **Glove** (300 d) and **Fasttext** (300 d);
- Contextual Embeddings: **BioMedNLP** (768 d);

4.Methodology

Training Methodology

- Train (**70%**), validation (**15%**) and test (**15 %**) sets;
- Gradient Clipping, Early Stopping, Dropout Regulation;



4.Methodology

Class Imbalance Handling :

- **SMOTE** (Synthetic Minority Over-sampling Technique)**;**
- **Class Weighting;**

Evaluation metrics : Balanced Accurcy, F1-score, Recall, Precision and AUC;

Tools : Python, Pytorch and WandB

5.Experiments & Key Results

Data Preparation for Machine Learning Models :

TF–IDF vectorization approach (with a maximum of **5000** features);

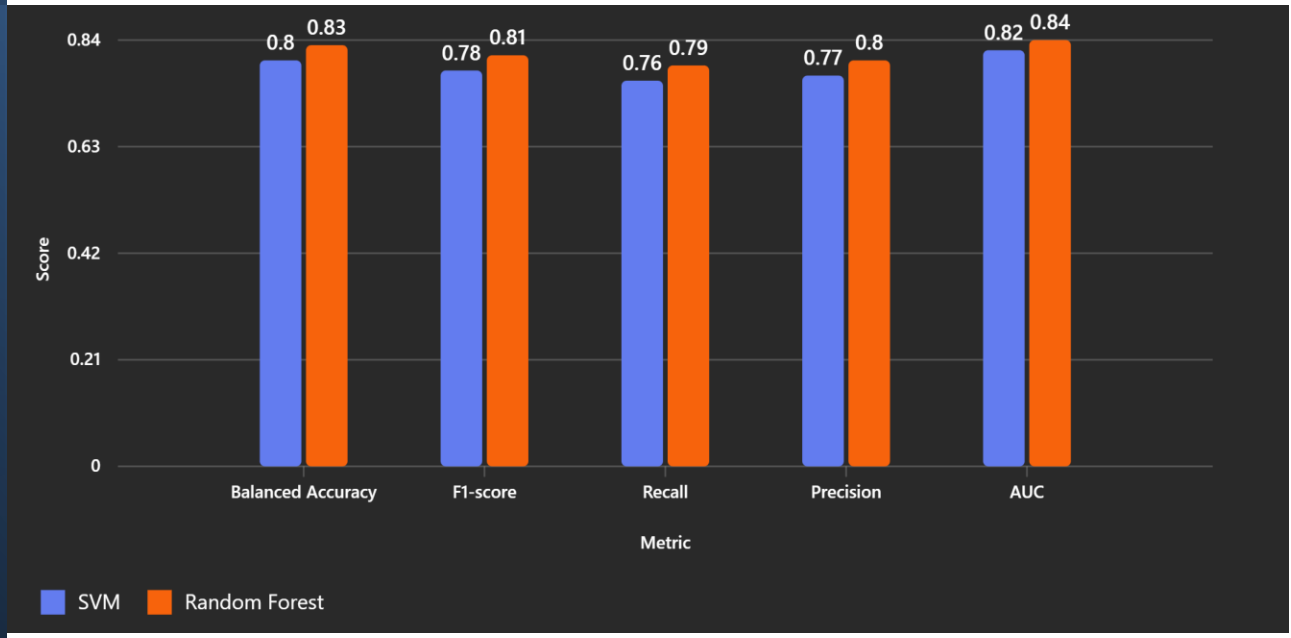
Data Preparation for Deep Learning Models :

- **MSKCC dataset** : Vocabulary size of **40,865** unique tokens (frequency lower than **3**);
- **Medical Abstract dataset** : Vocabulary size of **20,931** unique tokens (frequency lower than **2**).

5.Experiments & Key Results

Results on MSKCC dataset (Binary Classification Task)

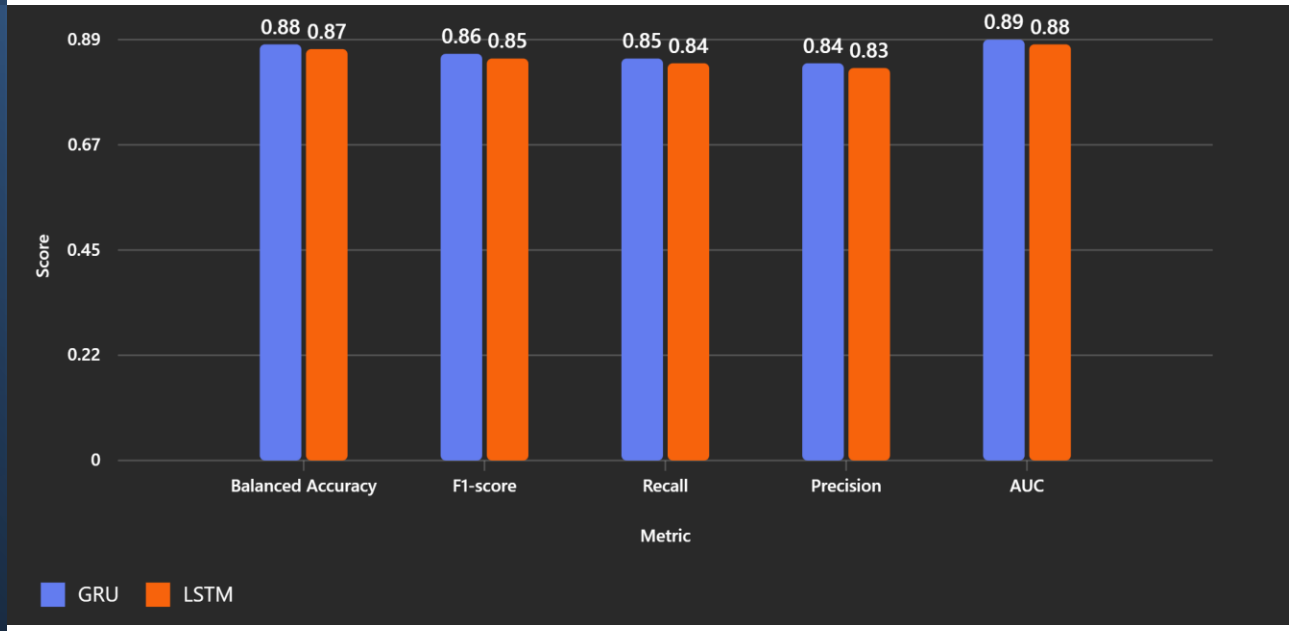
Machine Learning Models (with SMOTE applied)



5.Experiments & Key Results

Results on MSKCC dataset (Binary Classification Task)

Deep Learning Models (with Class Weighting)



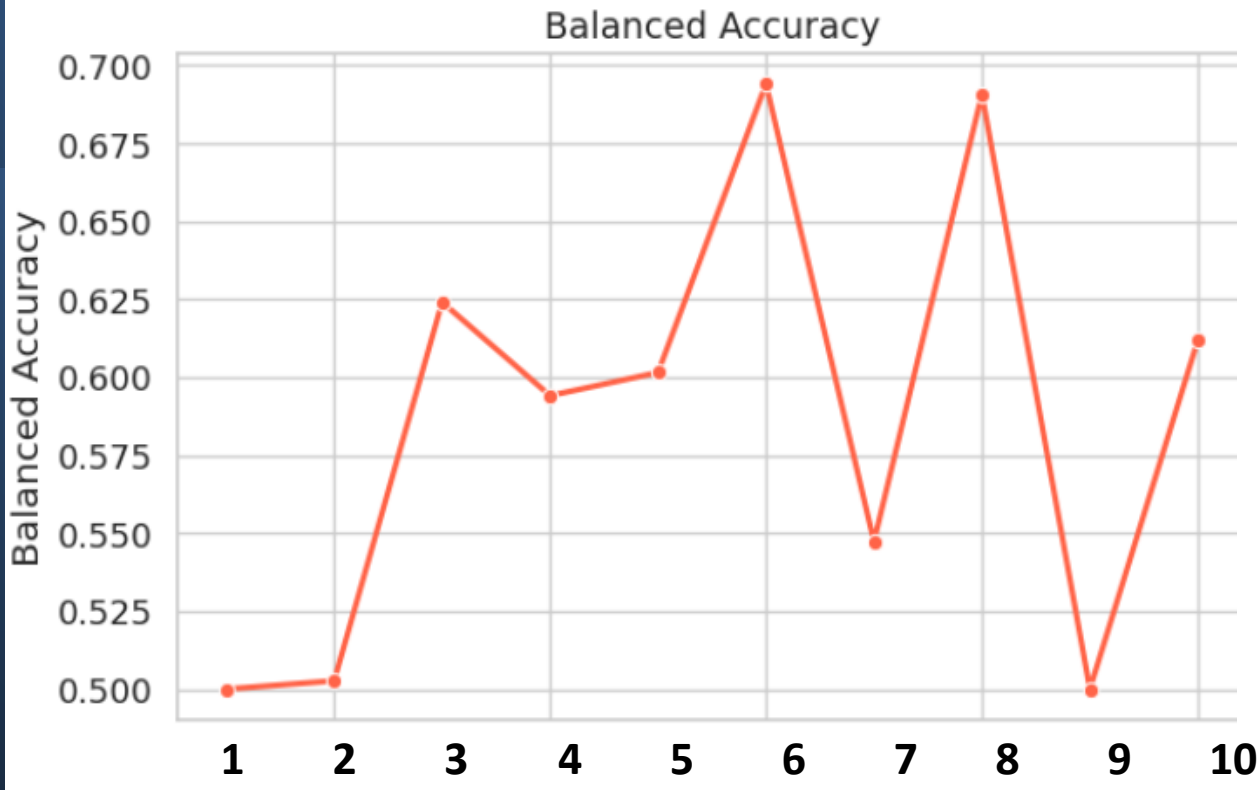
GRU/LSTM + Attention + BioMedNLP

GRU Training time : 1.65 (min/epoch);
LSTM Training time : 1.74 (min/epoch)

5.Experiments & Key Results

Few-Shot Learning

Results on MSKCC dataset (Binary Classification Task)

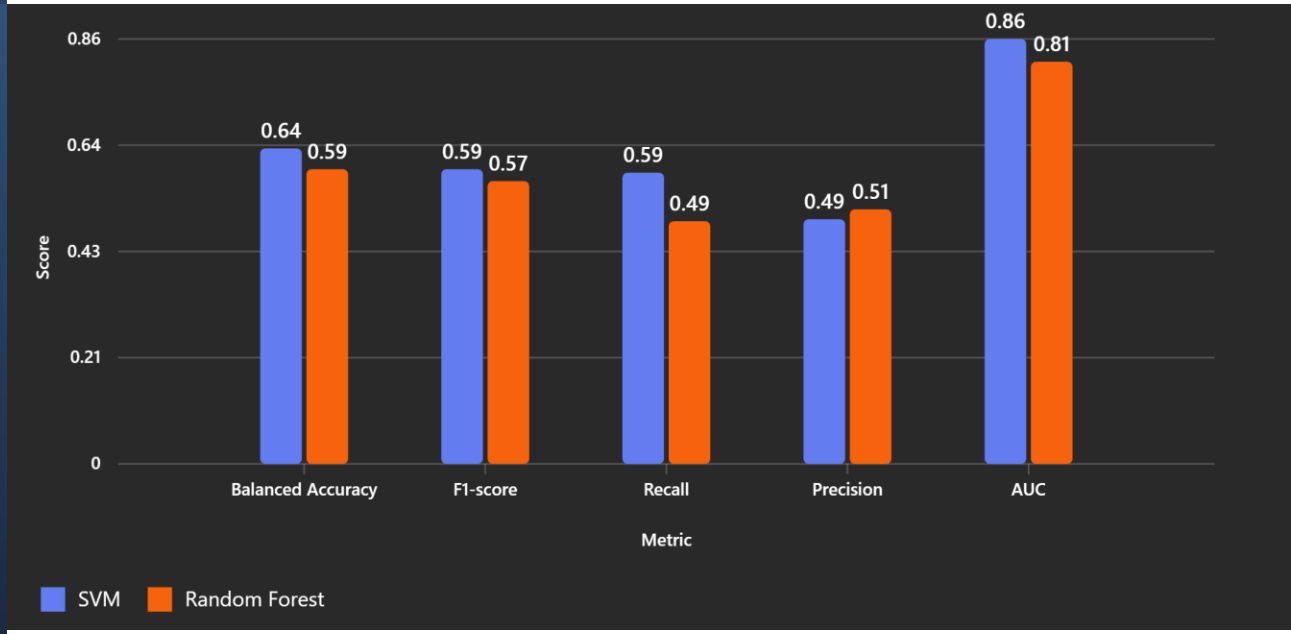


Bidirectional GRU model with Bahdanau attention using BioMedNLP

5.Experiments & Key Results

Results on Medical Abstract Dataset (Multi-Class Classification Task)

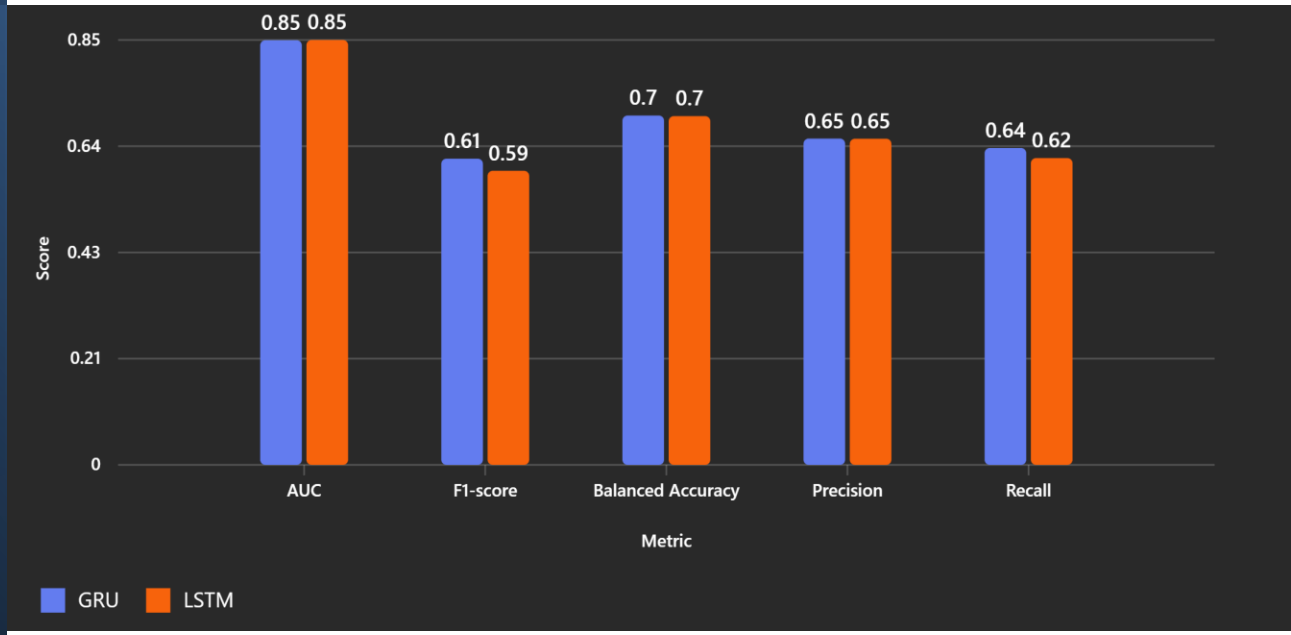
Machine Learning Models (with SMOTE applied)



5.Experiments & Key Results

Results on Medical Abstract Dataset (Multi-Class Classification Task)

Deep Learning Models (with Class Weighting)



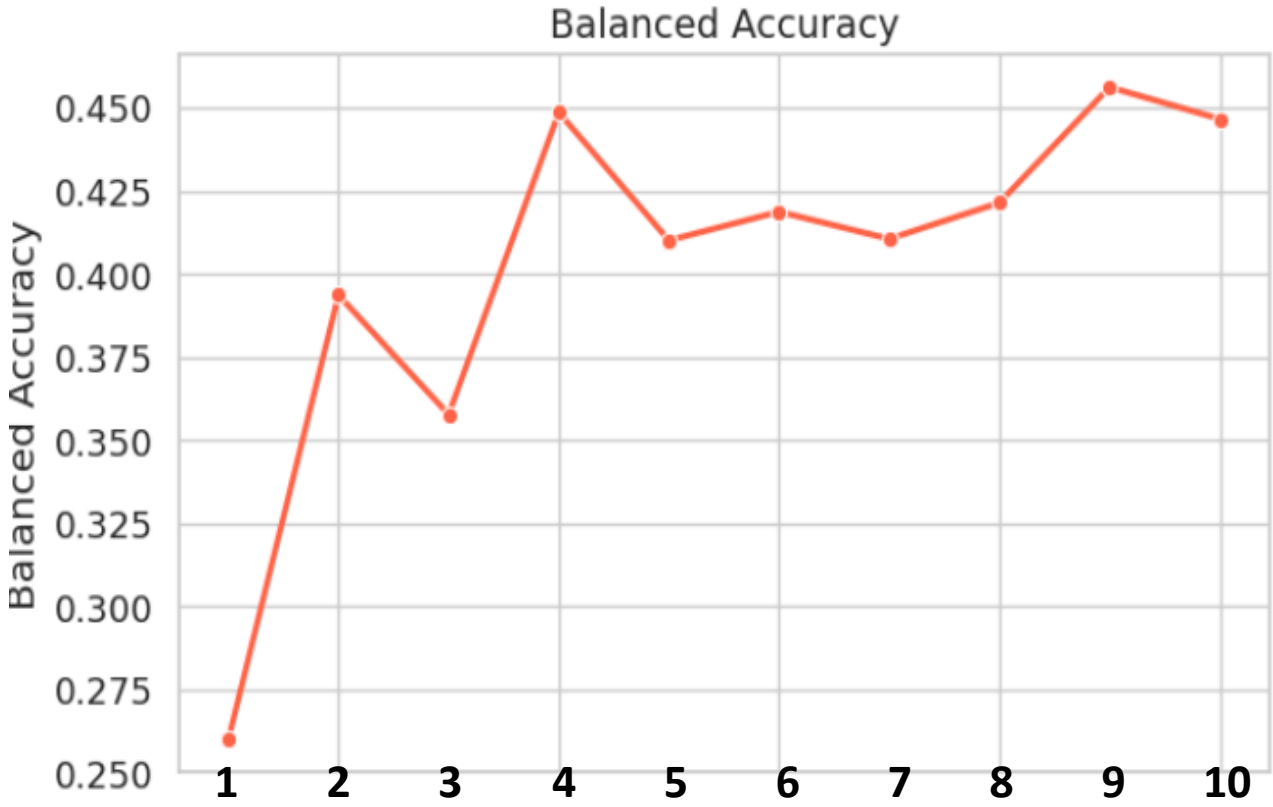
GRU/LSTM + Attention + BioMedNLP

GRU Training time : 5.49 (min/epoch);
LSTM Training time : 6.12 (min/epoch)

5.Experiments & Key Results

Few-Shot Learning

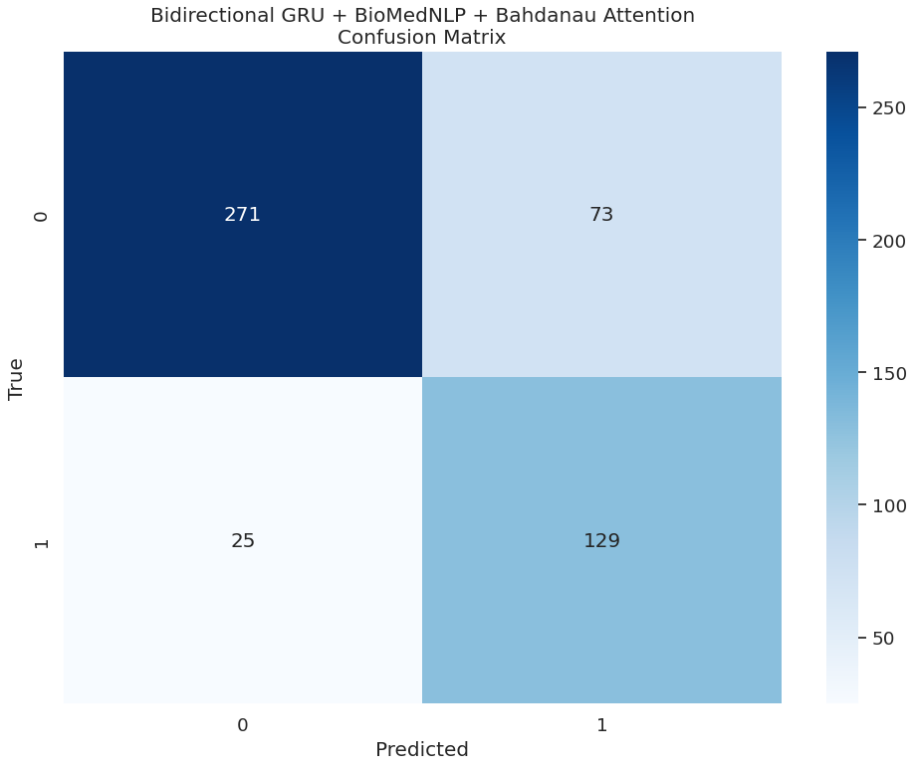
Results on Medical Abstract Dataset (Multi-Class Classification Task)

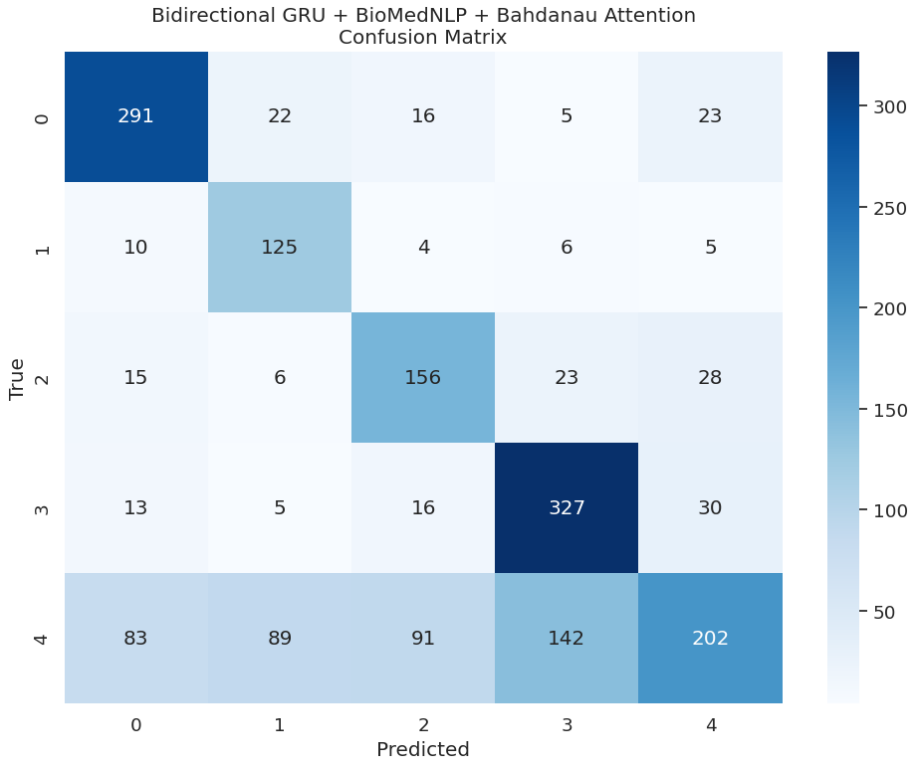


Bidirectional GRU model with Bahdanau attention using BioMedNLP

5.Experiments & Key Results

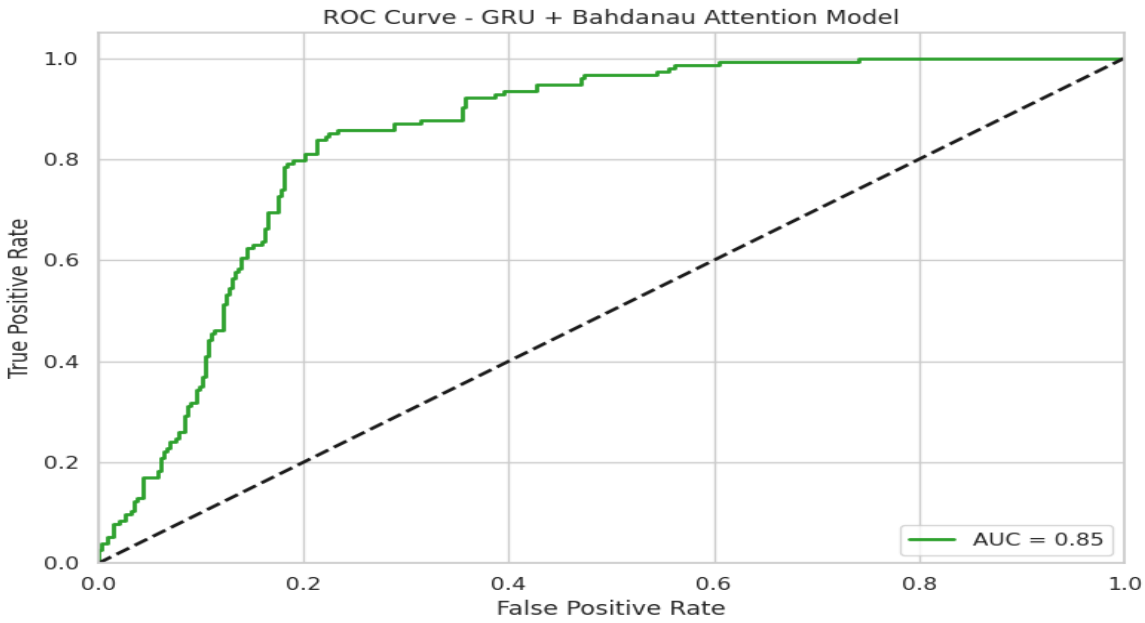
Bidirectional GRU + Attention + BioMedNLP

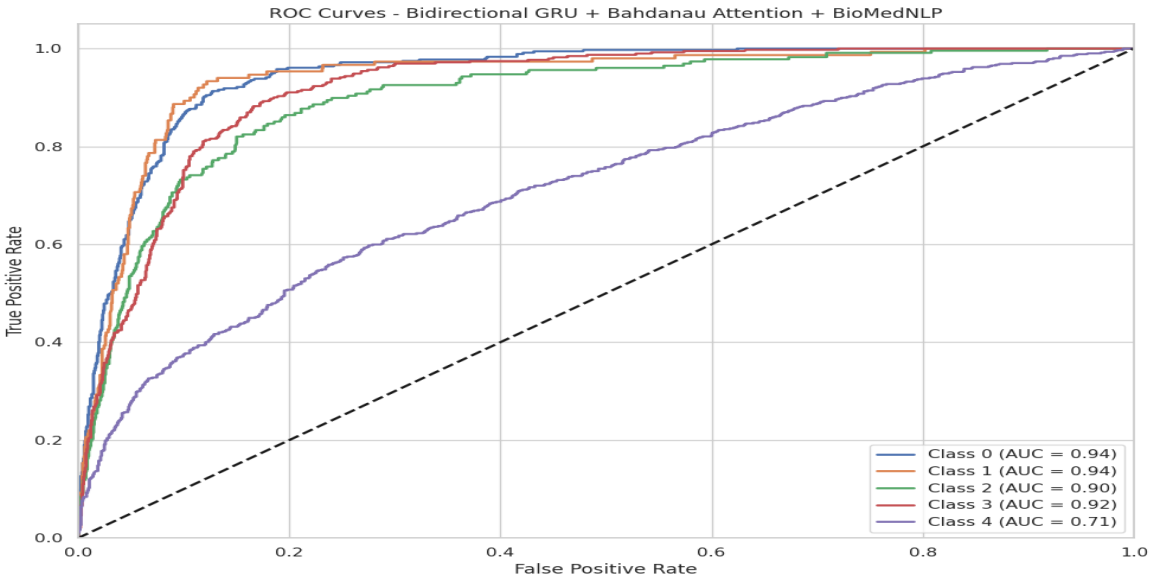




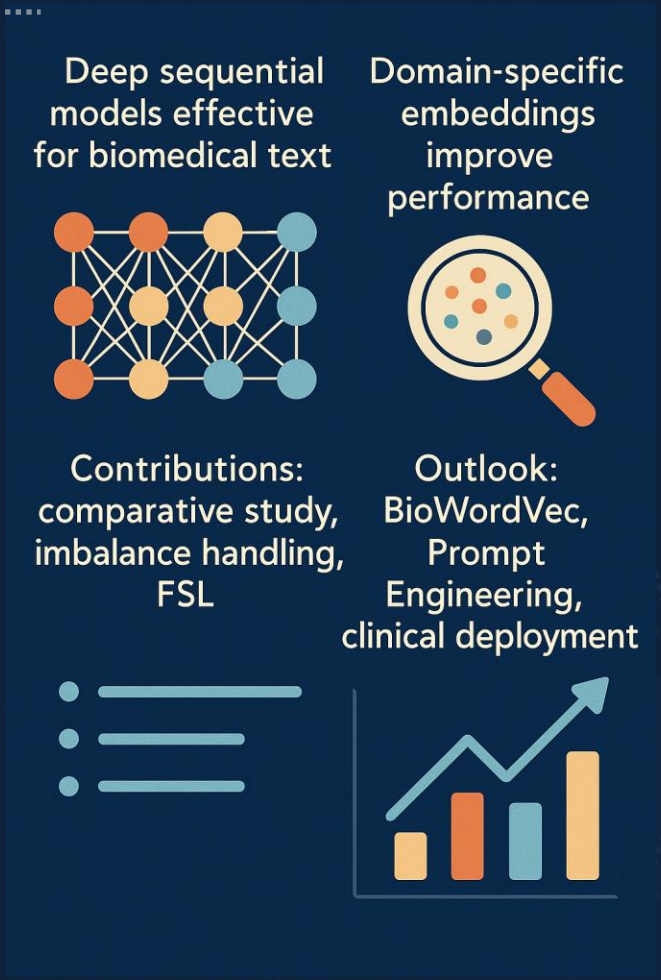
5.Experiments & Key Results

Bidirectional GRU + Attention + BioMedNLP



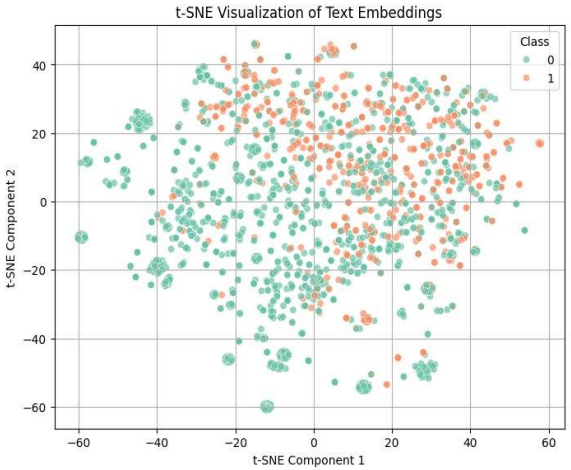


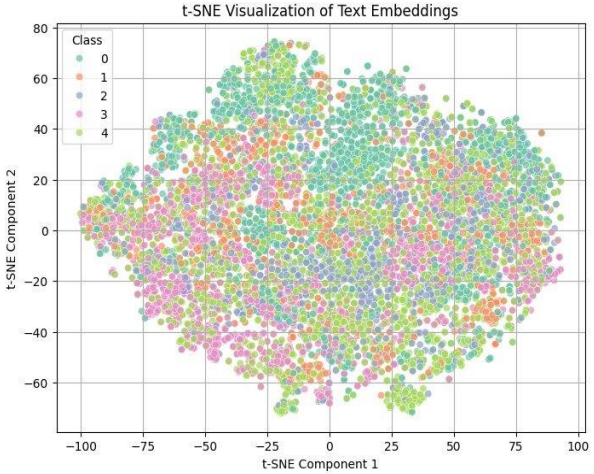
6.Conclusion & Perspectives

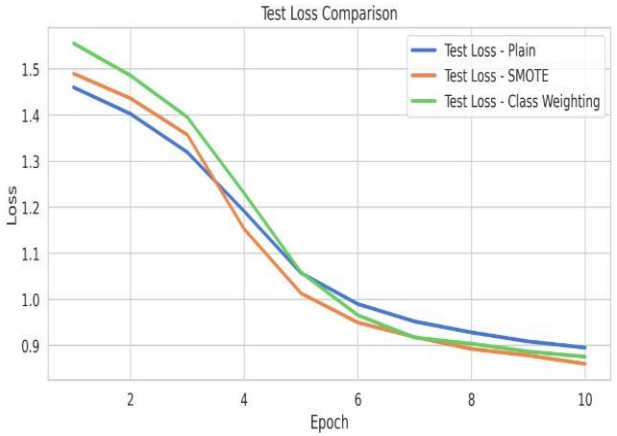


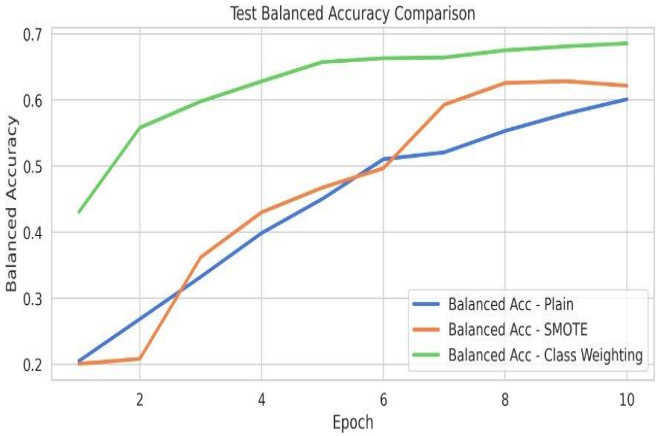
ANNEX

t-SNE visualization using BioMedNLP as embeddings



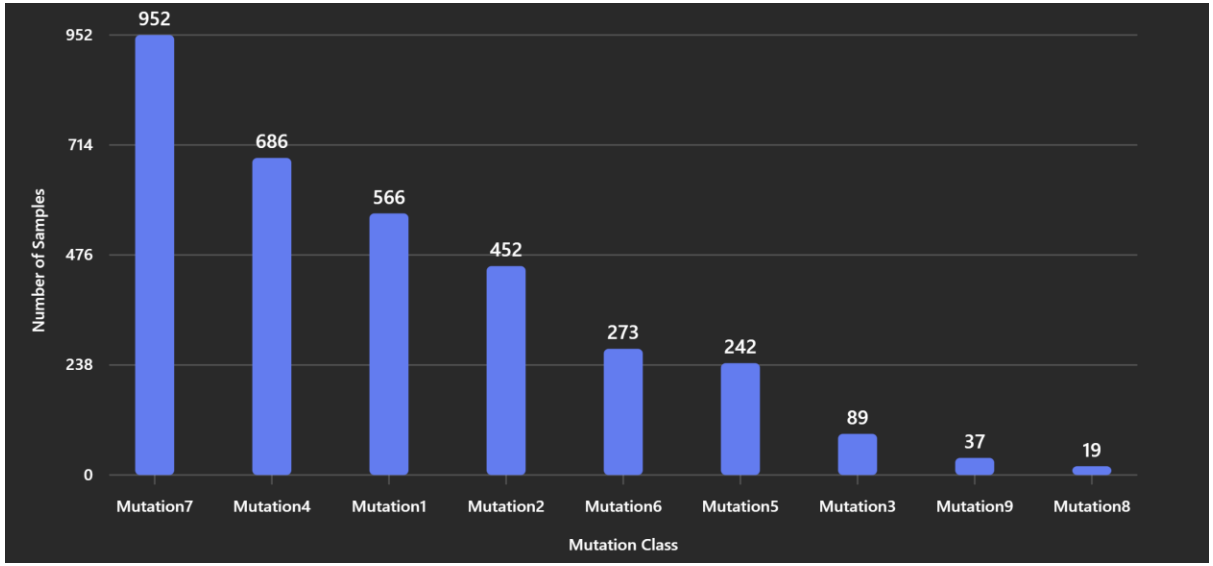




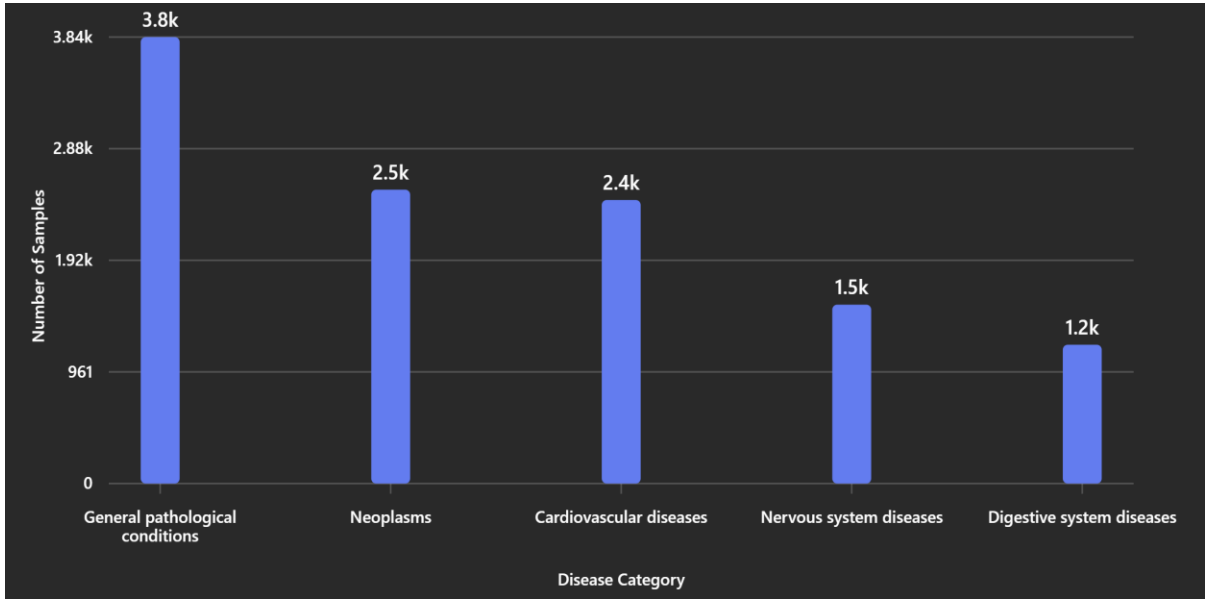


Bidirectional LSTM with BioMedNLP embeddings: Test Loss and Balanced Accuracy over epochs for different strategies (no resampling, class weighting, and SMOTE) on Medical Abstract dataset

ANNEX



Distribution of samples according to mutation classes in MSKCC dataset



Distribution of samples according to disease categories in the Medical Abstracts Dataset

Thank you!

Questions and discussion welcome.