**Supplementary file 1**

# Hybrid deep learning models for text-based identification of gene-disease associations

**Noor Fadhil Jumaa[1], Jafar Razmara[1*], Sepideh Parvizpour[2], Jaber Karimpour[1]**

[1]Department of Computer Science, Faculty of Mathematics, Statistics, and Computer Science, University of Tabriz, Tabriz, Iran

[2]Research Center for Pharmaceutical Nanotechnology, Biomedicine Institute, Tabriz University of Medical Sciences, Tabriz, Iran

**Table S1.** Summary of Studies on Machine Learning, Deep Learning, and Transformer-Based Methods for Gene-Disease Association Extraction.

| Study | Approach | Proposed Methods | Limitation |
|---|---|---|---|
| Bhasuran et al. (2018) | Machine Learning | Ensemble SVM with Word2Vec features | Struggles with accurately extracting associations from long or syntactically complex sentences. |
| Wu et al. (2019) | Deep Learning | RENET (CNN for representation + GRU/LSTM for relation extraction) | Needs improved named entity recognition (NER) and better identification of sentence boundaries to enhance relation extraction. |
| Bokharaeian et al. (2020) | Machine Learning | Linguistic-based model (negation signals and neutral candidates) | The limited availability of neutral candidates in the dataset reduces model robustness and generalizability. |
| Wang et al. (2020) | Machine Learning | HNEEM and HNEEM-PLUS (graph embedding + ensemble learning) | Model complexity leads to scalability challenges and a higher risk of overfitting on training data. |
| Lee et al. (2020) | Transformer-Based | BioBERT (biomedical pre-trained language model) | Requires extensive computational resources and long training times for pre-training large models. |
| Nourani et al. (2020) | Deep Learning | Deep-GDAE (CNN + BiLSTM with sentence-level attention) | Limited exploration of cross-sentence relationships in extracting complex associations. |
| Su et al. (2021) | Deep Learning | RENET2 (CNN + RNN with section filtering and ambiguity modeling) | Inefficient when extracting multiple relations, as the model processes only one gene-disease pair at a time. |

| | | | |
|---|---|---|---|
| Deng et al. (2021) | Transformer-Based | BioBERT (NER followed by relation extraction) | Model setup demands significant technical skills, and abstracts lacking explicit gene or disease are ignored, risking the loss of valuable data. |
| Xiang et al. (2022) | Machine Learning | HyMM (multiscale module decomposition + naïve Bayes integration) | Requires development of more adaptive and resilient methods for detecting biological network modules across varying data types. |
| Bokharaeian et al. (2023) | Deep Learning | CNN-LSTM, BERT-LSTM, PubMedBERT-LSTM models | Challenges remain in accurately processing complex linguistic patterns and varied sentence structures in biomedical text. |
| Dehghani et al. (2023) | Deep Learning | BioBERT-GRU (CNN + BiGRU on top of BioBERT embeddings) | Future improvements are needed to better capture ambiguous or indirect relationships through fuzzy relation modeling. |

**Table S2.** The EU-ADR dataset in summary

| Class | Number of unique genes | Number of unique diseases | Number of sentences |
|---|---|---|---|
| Positive | 150 | 95 | 213 |
| Negative | 16 | 9 | 19 |
| speculative | 24 | 20 | 30 |
| False | 73 | 40 | 93 |
| Total | 218 | 118 | 355 |

**Table S3**. The GAD dataset in summary

| Class | Number of unique genes | Number of unique diseases | Number of sentences |
|---|---|---|---|
| positive | 402 | 137 | 967 |
| negative | 544 | 209 | 1834 |
| No Semantic | 897 | 462 | 2529 |
| Total | 1131 | 535 | 5330 |

**Table S4.** The SNPPhenA dataset in summary

| | **Train** | **Test** | **Total** |
|---|---|---|---|
| Number of sentences | 935 | 365 | 1300 |
| Number of unique SNP (gene) | 287 | 130 | 417 |
| Number of unique Phenotypes (disease) | 268 | 92 | 360 |
| Positive | 702 | 170 | 872 |
| Negative | 91 | 29 | 120 |
| Neutral | 142 | 166 | 308 |

**Table S5.** Comparative summary of the number of instances per class in each dataset.

| Dataset | Positive Associations | Negative Associations | Neutral/Speculative/False Associations | Total |
|---|---|---|---|---|
| EU-ADR | 213 | 19 | 123 (Speculative + False) | 355 |
| GAD | 967 | 1834 | 2529 (No Semantic) | 5330 |
| SNPPhenA | 872 | 120 | 308 (Neutral) | 1300 |