Gaussian Distributed Prototypical Network for Few-shot Genomic Variant Detection

Jiarun Cao[1], Niels Peek[3,4], Andrew G Renehan[5,6], Sophia Ananiadou[1,2,*]
[1]National Centre for Text Mining, University of Manchester, UK
[2]The Alan Turing Institute, London, UK
[3]Centre for Health Informatics, University of Manchester, UK
[4]NIHR Biomedical Research Centre, University of Manchester, UK
[5]Division of Cancer Sciences, University of Manchester, UK
[6]Manchester Cancer Research Centre, University of Manchester, UK
{jiarun.cao, niels.peek, andrew.renehan, sophia.ananiadou}@manchester.ac.uk
(*) Corresponding author: Sophia Ananiadou

Abstract

Automatically identifying genetic mutations in the cancer literature using text mining technology has been an important way to study the vast amount of cancer medical literature. However, novel knowledge regarding the genetic variants proliferates rapidly, though current supervised learning models struggle with discovering these unknown entity types. Few-shot learning allows a model to perform effectively with great generalization on new entity types, which has not been explored in recognizing cancer mutation detection. This paper addresses cancer mutation detection tasks with few-shot learning paradigms. We propose GDPN framework, which models the label dependency from the training examples in the support set and approximates the transition scores via Gaussian distribution. The experiments on three benchmark cancer mutation datasets show the effectiveness of our proposed model.

Due to the ever-expanding biomedical literature, automated approaches in the biomedical text mining domain play an important role in mining gene interactions (Özgür et al., 2008; Trieu et al., 2020; Sahu et al., 2019), identifying biomarkers and exploring the genetic mutations, which can significantly reduce time and effort compared to traditional labour-intensive approaches. In particular, as a critical step in analysing the literature for cancer genomics data, text mining in cancer genomics studies (Birgmeier et al., 2020; Cejuela et al., 2017; Mahmood et al., 2016; Wei et al., 2013, 2018) has automatically identified novel somatic alterations such as single-nucleotide polymorphisms (SNPs), deletion and insertions, copy number aberrations, structural variants, and gene fusions.

For cancer genomics mutation extraction, the most representative works use either manually-crafted templates (Caporaso et al., 2007; Si and Roberts, 2018) or feature engineering (Cejuela et al., 2017; Wei et al., 2018) with machine learning-based approaches (Doughty et al., 2011; Wei et al., 2015; Si and Roberts, 2018). The main drawback of the traditional methods is that they are not competent with unseen categories. Nevertheless, as cancer research advances, thousands of new cancer genomes and exomes are identified and classified into new categories. There has been a lack of progress in automated genomic variant detection attempts. There may be a way around this problem by annotating more data for the model to capture new categories, but this would be highly costly in terms of time and labour costs in the cancer domain.

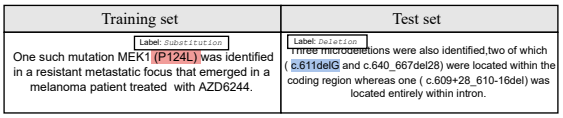

Figure 1: An example shows the semantic inconsistency issue between training set and test set. The entity 'P124L' from the training set differs substantially from 'c.611delG' in the test set, highlighting the challenge of predicting unseen categories by supervised learning-based models. This difference illustrates the difficulty traditional few-shot learning methods encounter when trying to recognize novel cancer genomic variants.

Intuitively, humans can understand a concept with a few samples, which drives the researcher to apply the few shot learning paradigm to downstream text mining tasks, such as named entity recognition (Cao et al., 2021; Settles, 2004), relation extraction (RE) (Yao et al., 2019; Zhou et al., 2014), and event extraction (EE) (Trieu et al., 2020; Björne and Salakoski, 2018). In FSL, a trained model rapidly learns a new concept from a few examples while retaining great generalisation from observed examples (Vinyals et al., 2016). Thus, if

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 26–36

we want to add a new category of genetic variants, we only need a few samples to activate the system without retraining the model. This way, we can dramatically lower the cost of annotation and training while retaining high-quality outcomes.

In a few shot learning iteration, the model is given a support set and a query instance. The support set consists of examples from a small group of categories. A model is required to predict the label of the query instance following the set of categories that appeared in the support set.

A conventional approach for identifying entities using Few-Shot Learning (FSL) involves decomposing this task into a sequence labeling problem, taking into account the label dependency between each token. To consider both the item similarity and label dependency, previous attempts have utilized Conditional Random Fields (CRFs) in few-shot learning sequence labeling schemes (Hou et al., 2020; Das et al., 2021; Wang et al., 2022; Fritzler et al., 2019; Yang and Katiyar, 2020; Wang et al., 2021; Li et al., 2020). However, learning the scoring and transition scores of CRF presents distinct challenges in the few-shot situation.

Regarding the scoring score, prior works (Hou et al., 2020; Das et al., 2021; Wang et al., 2022; Fritzler et al., 2019) relied on the Prototypical Network (Snell et al., 2017) to average the embeddings of each label's support instances as label representations. These frequently distribute densely in the embedding space, often leading to mistaken predictions. Learning the transition score using only a few labeled data also poses challenges (Yang and Katiyar, 2020; Wang et al., 2021; Li et al., 2020), as the prior label reliance in the source domain cannot be directly transferred and leveraged due to the difference in the label set.

Figure 1 exemplifies these issues. The semantic differences between entities in the training and test sets result in difficulties for supervised learning-based models in predicting unseen categories. For instance, in the context of recognizing cancer genomic variants, these models often fail to identify entities in novel categories, as these categories usually contain distinct entity mentions from known entities.

In this paper, we propose a novel sequence labelling method to alleviate this problem in the settings of few shot learning. Specifically, we proposed a Gaussian Distributed Prototypical Network(GDPN) which has two merits: (1) we propose an interactive prototype network to capture the interaction from a few samples between different categories in the support set, and utilise those prototypical representations to provide more training signals towards the scoring function for the CRF-based model. (2) The conventional CRF models require a large amount of data sample to gain a precise estimation, which could suffer from the huge gap between a few samples in a specific label, we utilise Gaussian distribution to estimate the transition scores to avoid the randomness caused by scarce samples.

We experiment with the proposed models on the different benchmark cancer genomic dataset (Lee et al., 2016; Wei et al., 2013; Doughty et al., 2011). The experiments show that our methods can improve the performance of the genomic variant detection with the FSL settings. To summarise, our contributions to this work include:

- We formulate cancer genomic variant detection as a few-shot learning problem to extend this task to novel mutation types and provide a baseline for this new research direction. To our best knowledge, this is a new branch of research that has not been explored on this subject.
- We propose a novel method, namely GDPN, which models the specific label dependency and overcomes the data fluctuation in the few-shot learning setting.
- Experimental results show that our proposed model outperforms other competitive FSL baselines and the state-of-the-art CRF-based baselines on the benchmark datasets in cancer genomic domain. Further analyses demonstrate the model's effectiveness.

1 Formulation

Our goal in this work is to formulate cancer genomic variant detection as a FSL problem, which has not been done in prior work. To achieve this, we first present the FSL framework and specify symbols and terminology in this section 1. Then we illustrate the proposed method in the following section 2.

1.1 Few Shot Learning

In Few shot learning (FSL), we preliminarily assign two sets: support set $\mathcal{S}$, which contains classified

samples, and query set $\mathcal{Q}$, which contains unclassified samples. Models can predict the label of an instance x from a query set $\mathcal{Q}$, by learning from a support set $\mathcal{S}$ and a label set C. Prior FSL investigations used an N-way K-shot configuration with N clusters representing N categories and K data samples.

Since we cast this task as a sequence labelling problem and adopt BIO tagging schema (B represents beginning of an entity, I represents intermediate of an entity, O represents outside an entity), we extend the N-way K-shot to $2*N+1$ way K shot, where 2*N clusters denote the B and I categories, and 1 cluster denotes O label.

Therefore, given a word sequence $X = \{x_1, x_2, \ldots, x_n\}$ and its corresponding label sequence $Y = \{y_1, y_2, \ldots, y_n\}$, the support $\mathcal{S}$ can be represented as:

$$S = \{(X_0, Y_0), (X_1, Y_1), \ldots, | \ (X_{(2*N+1)*K}, Y_{(2*N+1)*K})\} \tag{1}$$

Where $(2 * N + 1) * K$ is the total number of samples in the support set $\mathcal{S}$.

1.2 Linear Chain CRF

Conditional Random Fields (CRFs) (Wallach, 2004) are undirected statistical graphical models, which are well suited to tackle sequence labelling problem. Following on the above section, given a word sequence $X = \{x_1, x_2, \ldots, x_n\}$ and its corresponding label sequence $Y = \{y_1, y_2, \ldots, y_n\}$, linear-chain CRFs define the conditional probability of a label sequence given an input sequence to be:

$$P(Y|X) = \frac{\exp\left(\sum_{k=1}^{n} U(x_k, y_k) + \sum_{k=1}^{n-1} T(y_k, y_{k+1})\right)}{Z(X)} \tag{2}$$

where $Z(X)$ is a normalization factor of all state sequences. Note that $U(.)$ is the scoring function that calculates the probabilistic score of label y for each token in the sequence X. $T(.)$ is the transition function that calculates the transit score between the adjacent labels y_k and y_{k+1}.

2 Method

2.1 Instance Encoder

We first map discrete words to a continuous high-dimensional vector space to simplify neural network training using BioBERT (Lee et al., 2020), which is a pre-trained biomedical language representation model and had shown great effectiveness on many downstream biomedical text mining tasks. Formally, Given the token sequence $x = \{x_1, x_2, \ldots, x_n\}$, we have:

$$H = h_1, h_2, \ldots, h_n = BioBERT(x_1, x_2, \ldots, x_n) \tag{3}$$

2.2 Interactive Prototype Encoder

This module generates a representative vector for each label t in the support set S from the overall representations of its instances. Instead of employing the original Prototypical Network suggested in (Snell et al., 2017), which determines all representation vectors equally, we claimed that the supporting vectors are conditionally important with respect to each query $q \in \mathcal{Q}$, therefore, model the interactions from each label. Formally, to compute the prototype for a class $t \in T$, it collects all of the instance's representations and calculates them as the supporting vectors' weighted sum. The weights are determined by the attention mechanism in accordance with the query representation:

$$a_t^j = \sum \sigma(f(H_t^j) \odot f(q)) \tag{4}$$

$$\alpha_t^j = \frac{exp(a_t^j)}{\sum_{H^k \in \mathcal{S}} exp(a_t^k)} \tag{5}$$

$$Pr_t = \sum_{H^t \in \mathcal{S}} \alpha_t^j f(H_j) \tag{6}$$

where Pr_t denotes the Prototypical representation of label $t \in T$, $\odot$ denotes element-wise product. f represents the encoding function and is $BioBERT$ in our paper.

2.3 Gaussian Distributed Prototypical CRF

In section 1.2, we already learn that CRF layer consists of a scoring function $U(.)$ and a transition function $T(.)$. We compute these two components separately. The scoring function represents a value U for the label y given our token x_i vector at the i-th timestep. Prior works leverage the output of LSTM as the U, where it is the so-called LSTM+CRF framework (Huang et al., 2015) that has been applied to most of the conventional NER tasks.

Instead of using the output of LSTM to gain the output U from scoring function, we first calculate

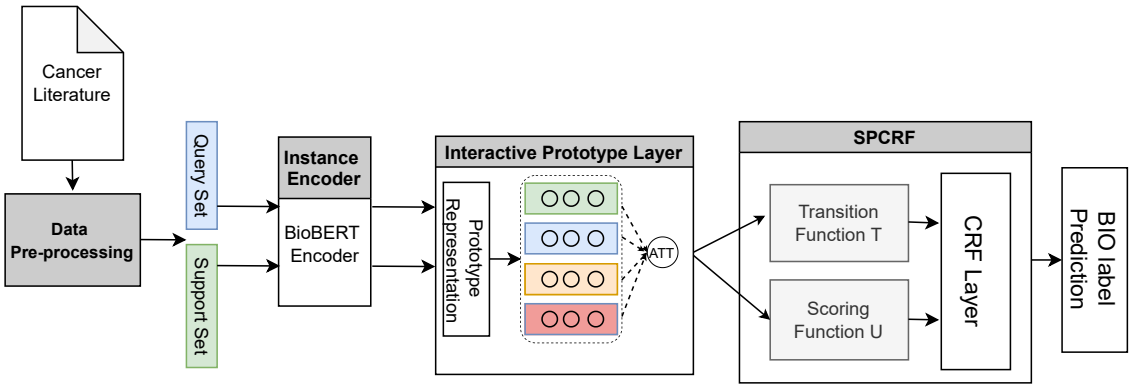

Figure 2: Model architecture. ATT denotes weighted sum attention operation.

the correlation between each token x_i representation and the prototype representation Pr_t for label y_t.

$$s_i = f_S(y_t, x_i, \mathcal{S}) = Pr_t \odot h_i \tag{7}$$

Accordingly, the scoring function $U()$ for the entire sequence is gained via the sum of each token score as follows s_i :

$$U(y, x, \mathcal{S}) = \sum_{t=1}^{n} f_S(y_t, h_t, \mathbf{S}) \tag{8}$$

In terms of the transition score, the conventional CRF model optimizes the transition function $T(.)$ from massive data samples, which overcomes the data fluctuation problem to a large extent. However, the few data samples can achieve dramatic data randomness. The transit function lacks the optimization process, thus resulting in huge data bias representing the probability of transition of two adjacent labels. To alleviate this problem and smoothen the randomness caused by a few samples, we adopt Gaussian distribution as our transition function and utilize mean value μ and variance value σ to approximate the transition score as follows:

$$\mu_{ij} = W_\mu(Pr_i; Pr_j) + b_\mu \tag{9}$$

$$\sigma_{ij} = \exp\left(W_\sigma(Pr_i; Pr_j) + b_\sigma\right) \tag{10}$$

Where ; denotes concatenation operations. Like linear chain CRF, our transition function for the entire sequence is achieved as follows:

$$T(y) = \sum_{i=1}^{n-1} T(y_i, y_{i+1})) \tag{11}$$

Therefore, the probability of label sequence Y given the token sequence is as same as the conventional CRF model:

$$P(Y|X) = \frac{\exp\left(\begin{array}{c}\sum_{k=1}^{n} U(x_k, y_k) \\ + \sum_{k=1}^{n-1} T(y_k, y_{k+1})\end{array}\right)}{Z(X)} \tag{12}$$

Where and Z(x) is normalization factor in order to get a probability distribution over sequences. In the inference stage, we use Viterbi algorithm (Forney, 1973) as with traditional CRF to find the optimal path from the input.

3 Experiment

3.1 Dataset

In this work, we implement the proposed method on three benchmark datasets. The relevant statistical figures have been listed in Table 1. TmVar is a sequence variant corpus derived from Pubmed abstracts, which contains a large number of sequence variants at both the protein and gene level using a standard nomenclature for sequence variants created by the human genome variation society (Wei et al., 2013). TmVar includes 500 PubMed abstracts and titles with 871 variants.

BRONCO (Lee et al., 2016) is now the most extensive full-text cancer variant corpus annotated with information about genes, diseases, medicines, and cell lines associated with the variants. BRONCO has 108 full-text papers with 403 gene variations, as indicated in Table 1. EMU (Doughty et al., 2011) searched mutations, gene mentions, and disease connections by retrieving a set of PubMed abstracts that were possibly beneficial for finding mutations. EMU contains two subsets which are Breast cancer and prostate

cancer, respectively. As we can see from Table 1, EMU consists of 109 PubMed abstracts with 172 variants.

3.2 Data Pre-processing

FSL does not allow us to directly use the dataset's splits since the label types in the training set and the testing set are not congruent. As a result, we adopted the scheme conducted in (Lai et al., 2020) and have further divided these datasets to meet three requirements for FSL:

- Label types in the train set are distinct from those in the testing and development sets. In another word, there is no overlap regarding the label types between training/development/testing sets.
- The label type contain less than 5 samples are abandoned.
- The training set should contain as many samples as possible.

We re-split the dataset based on the standards above. As the label types of EMU and BRONCO are quite limited to support the FSL setup, we combine both dataset as a whole to underpin the FSL training and testing process. The final splits are shown in Table 2.

3.3 Implementation Details

We adopt a mini-batch mechanism to train our model, with a batch size of 2 and a learning rate of 1e-5. A warm-up strategy and dropout with 0.1 probability are introduced to prevent the model from over-fitting. All parameters are optimized using Adam (Kingma and Ba, 2014). Furthermore, we also adopt an episodic training scheme that has been commonly adopted in fsl, and we used the sample evaluation methods in (Cong et al., 2020); an entity is counted as correct only if its label and its textual span are both correct.

3.4 Baseline Models

Since the scope of our task is NER with fsl settings, we compare the proposed model with two types of baselines: the state-of-the-art FSL models that have been applied in many areas and the typical NER models commonly used for NER tasks. For FSL baseline models, we applied 5 well-adopted ones which include (1) Matching Network (Vinyals et al., 2016) adopted cosine similarity as a prototypical score with the averaging operation. (2) Proto Network (Snell et al., 2017) used Euclidean Distance as the similarity metric with the averaging prototype. (3) Proto+Dot (Lai et al., 2020) used a dot product to compute the similarity. (4) Proto+Att (Lai et al., 2020) used a weighted sum prototype with Euclidean Distance. (5) Relation (Sung et al., 2018) builds a trainable distance function and a neural network to measure the similarity.

In terms of the CRF-based baselines, they can be divided into two groups: The first group consists of vanilla CRF sequence labeling models: (1)BiLSTM+CRF (Luo et al., 2018) utilizes the BiLSTM layer to map the semantics features to a higher dimension and CRF layer is to model the label's consistency. (2) BERT+CRF (Dai et al., 2019) is similar to BiLSTM+CRF instead of using BERT for feature extraction. The second group consists of the state-of-the-art CRF sequence labeling models for FSL NER tasks [1]: (1) CONTAINER (Das et al., 2021) optimized a generalized objective of differentiating between token categories based on their Gaussian-distributed embeddings. This effectively alleviates overfitting issues originating from training domains. (2) FEW-NERD (Ding et al., 2021) released a massive-scale FSL NER dataset and proposed the corresponding baseline models that combined BERT tagger with Prototype network. (3) Decomposed Meta-Learning (Ma et al., 2022) took the few-shot span detection as a sequence labeling problem and trained the span detector by introducing the model-agnostic meta-learning (MAML) algorithm to find a good model parameter initialization that could fast adapt to new entity classes.

4 Results

Table 4 and Table 5 in Appendix sector show the precision, recall, and F1 score of the baseline models and the proposed model on the three benchmark datasets under N-way K-shot few-shot learning settings. Unlike the conventional few shot learning tasks using 5 or 10 ways and shots for the settings, we utilize 1-to-3 ways and 1-to-5 shots due to the limited scale of the datasets. Additionally, we also evaluate the model by test epoch, which relates to the number of samples included in the test set, to verify the effect of the data fluctuation on the model's performance.

[1]We only adopt the baseline models that are applicable for our datasets and have the same settings with the proposed model.

	TmVar	EMU	BRONCO
Corpus Type	title and abstract	title and abstract	full-text
Number of Documents	500	109	108
Cancer Variants	871	172	275
Mutation Types	Sub, Del, Ins, Dup, InDel, SNP, FS	Sub, Del, Ins, SNP, FS	Sub, Del, Ins, InDel, SNP, FS

Table 1: Statistics of the evaluated datasets

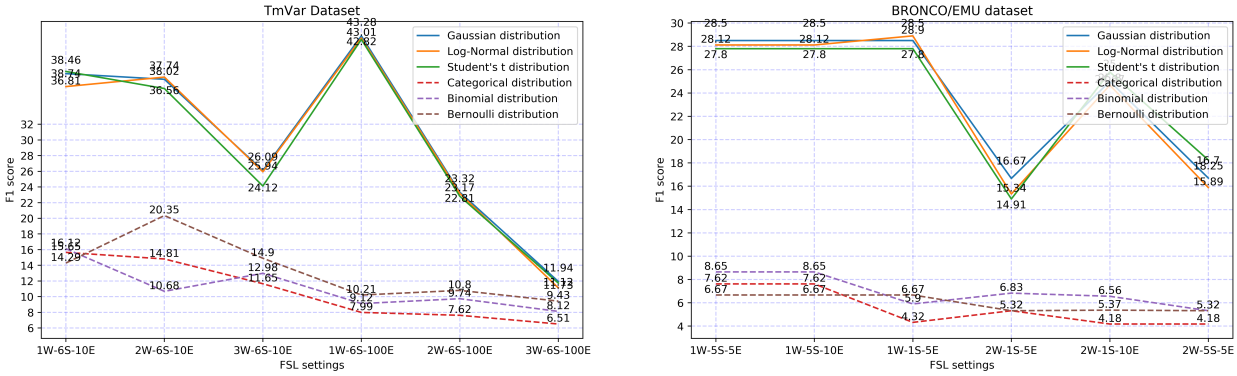

Figure 3: Experimental results with various of distributions, W, S, E denote Ways, Shots, and Epochs, respectively.

	training set	test set
TmVar	SUB, DEL, SNP	INS, FS, DUP
BRONCO & EMU	SUB, SNP	DEL, FS

Table 2: Label Splits for three datasets.

4.1 From the perspective of FSL Settings

We first evaluate the results from the perspective of different test settings. To be more concrete, we test the effect of N-way, K-shot, and test epoch, respectively. We can see from both Table 4 and Table 5 in Appendix the performance of the models on 1-way K-shot is always better than 2-way and 3-way K-shot. Statistically, the vanilla NER models drop 0.7% and 4.5% on average from 1 way to 2-way and 3-way given a certain number of shot, the general FSL models drop 3.12% and 4.86%, while CRF-based models drop 3.54% 6.13% and under the same circumstance. The reason is the fact that the increase in the number of classes leads to a larger scope of the probability distribution, resulting in the lower results.

The above demonstrates the influence of the N-way. Next, we analyze the effect of K-shot. As shown in Table 5, the baseline and proposed models mainly achieve better performance while K increases with a fixed N-way. In the FSL models in general, comparing the performance from K = 1 to K = 5 given the 1-way setting, the prototype-based models boost 4.94% F1 score on average, the other FSL models, i.e., Relation network and Matching network increase 3.58% F1 score on average, and even Vanilla NER model improves 5.24% on average. In the CRF-based models, we can also notice a 3.79% F1 increase. This unified tendency indicates that the added shots are able to benefit the models to gain more semantic features given a certain label, which is consistent with the experimental results we can observe from the other works (Lai et al., 2020; Das et al., 2021; Ma et al., 2022).

We also evaluate how the number of test epochs affects the results. We initially speculated that the test epoch determines the number of samples involved in the test loop, reflecting the influence of data randomness on models' performance. Thus, the lower test epoch should achieve higher performance improvements on a specific model, as the data randomness issue is more severe when the number of the data sample is smaller. However, the results suggest that different settings of the test epoch do not straightforwardly relate to the data randomness. As noticed in Table 4 and Table 5 in Appendix, the model's performance can be either higher or lower with different test epochs. When we keep the N-way and K-shot fixed, the test epoch cannot unveil the data fluctuation issue.

4.2 From the Perspective of Models

Then, we analyze the experimental results from the perspective of the model types. We can notice that in both Table 4 and Table 5 that general FSL

models outperform the vanilla NER models to a large extent. The reason is that the vanilla NER models struggle with the insufficient semantic features for each label, thus resulting in an unqualified transition matrix to model the label's consistency. BERT+CRF model exacerbates this trend due to the specific tokenization approach, WordPiece (Song et al., 2020), which is a more fine-grained way to split the words into subwords.

On the other hand, in the general FSL models, prototype-based (Proto, Proto+Att, Proto+Dot) models outperform the Matching network and the Relation network in all FSL configurations. Proto+Att and Proto-Dot are marginally better than Proto among prototype network models, with an average performance improvement of 2.18% and 1.96% F1 scores on the three benchmark datasets. The reason can be inferred that the interactive information amongst each label is integrated by Att and Dot operations, which naturally gains more benefits from the data samples. The proposed model outperforms the prototype-based models with an average 10.33% F1 score gap on BRONCO/EMU dataset and an average 10.99% F1 score gap on TmVar dataset.

Compared to the CRF-based models, our model is also built upon the CRF architecture, which leverages the label's dependency to cast this task as a sequence labeling problem. As we can notice in Table 4, our model outperforms the baseline models under different settings, and achieves 1.39% F1 score advance in TmVar dataset compared to each state-of-the-art results. For BRONCO/EMU dataset, we can notice that our model achieves the competitive results. When there is under the settings of 1Way-5Shot-5Epoch, 1Way-1Shot-5Epoch and 2Way-1Shot-10Epoch, our models outperform all the baseline models. The model gains these improvements due to the fact that successfully reducing the illegal label transition from CRF-based models. Our proposed model approximates the transition scores via prototypical representations, and optimize it by Guassian distribution to alleviate the huge data fluctuation issue caused by limited number of training samples.

4.3 Effectiveness Analysis

4.3.1 Ablation Study

We evaluate the model components in three aspects shown in Table 3. Instead of using Gaussian distribution, we generate the transit score directly, the model performance drops 0.38% and 0.61% F1 score, respectively, on TmVar and BRONCO/EMU datasets. It indicates that our Gaussian distribution estimation can alleviate the data uncertainty to some extent and thus estimate a more accurate transit score to reflect the data samples. Furthermore, we also replace the interactive prototype layer with the vanilla prototype network, and we can notice the model performance decreases with a 3.98% and 3.81% F1 gap. We can infer that the interactive prototype layer can integrate with different categories by giving different weights to the prototypical representations. Finally, we changed our BioBERT instance encoder to raw word2vec embedding (McCormick, 2016). The results dropped 1.35% and 0.38% F1 scores on the datasets, which shows the effectiveness of the BioBERT in encoding the semantic information.

5 Discussion

As shown in Figure 4, we utilize two cases to demonstrate the effectiveness of the proposed model. Specifically, we compare the proposed model with the FSL and conventional NER models, respectively, to showcase our model's advance.

The upper figure is a comparison between the proposed model and a conventional CRF-based NER model. We can notice that our model correctly assigns the "E746-A750" a label "B-DEL", while BiLSTM-CRF model wrongly predicts it as "O." As the label "B-DEL" rarely appears in the training set, BiLSTM-CRF model struggles with capturing its relevant semantics and assigning "B-DEL" to the correct token spans. Our model can predict the "B-DEL" for the token "E746-A750" credited to the prototype representation of the label "B-DEL" that has been learned in the support set.

The following case compares the proposed model and a prototype-based FSL model. Our model successfully predicts "c.370-371insA" as "B-DEL" while Prototype Network predicts its as "I-DEL". This is due to the fact that Prototype Network only learns the transition scores via prototype representations of specific categories. Although the prototype representation provides the feature of label "DEL" to some extent, the model still miscognizes it as "I-DEL" because of the data randomness. Our model overcomes this issue according to suppress the data fluctuation via Gussinan distribution, therefore predicting this case correctly.

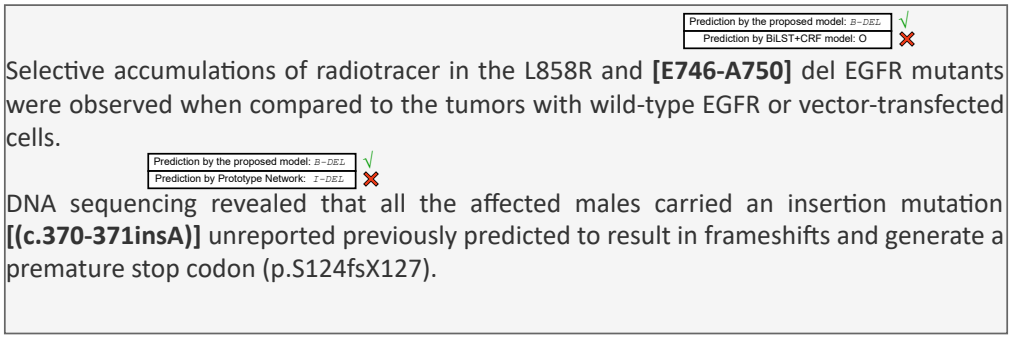

Figure 4: Two cases of the prediction between the proposed model and baseline models.

	TmVar			BRONCO & EMU		
	Precision	Recall	F1-score	Precision	Recall	F1-score
Raw model	32.20%	66.00%	43.28%	50.00%	20.00%	28.57%
(-) Gaussian Distribution	31.82%	65.84%	42.90%	46.67%	20.00%	27.96%
(-) Interactive Prototypical Representation	28.14%	63.10%	38.92%	36.67%	18.00%	24.15%
(-) BioBERT	26.84%	62.58%	37.57%	34.98%	18.00%	23.77%

Table 3: Experimental results of different baseline models on EMU PCa and BCa datasets.

5.1 Empirical Experiment of Distributions

Gaussian distribution is leveraged to estimate the transition scores to smoothen the fluctuation caused by scarce samples. In this section, we also conduct an empirical experiments to test the model's performance with different distributions [2]. The distributions can be divided into two groups, the first group is discrete variable distributions including Categorical distribution, Binomial distribution, and Bernoulli distribution. These group of distributions gain much lower results shown in Figure 3, because the range of the distribution function is discrete, and the output of possible values is finite. The second group is continuous variable distributions, including Gaussian distribution, Log-Normal distribution and Student's t distribution. As we can see from Figure 3, although these distributions turning out to be slightly higher or lower in a limited range, Gaussian distribution still achieves the best results in majority of the settings. We speculate the reason is that Gaussian distribution can better eliminate the influence of outliers in few sample scene, so as to accurately grasp the central tendency and discrete trend of data, therefore we empirically apply Gaussian distribution to our method.

5.2 Error Analysis

We also conducted the error analysis of predictions to demonstrate the models' bottleneck. 81.7% of them attribute to long-span errors, which means our model is relatively weak in predicting the textual spans that constitute more than one token. By 'long-span errors', we refer to instances when our model only predicted a portion of the total relevant token span, or when our model failed to properly identify and predict an entity that spans multiple tokens. This does not necessarily indicate a deficiency with the tokenization process. In fact, many surface forms of mutation events do consist of more than one token. However, the model struggles to capture these instances consistently, leading to these "long-span errors". This challenge appears to be a common issue in few-shot learning for this type of NER task, which often require more comprehensive training to effectively capture and predict entities that consist of multiple tokens.

On the other hand, 9.8% of errors are because our model does not recognize the target entities, thus just assigning them a "O" label. Finally, 8.5% of errors can be summarized that our model successfully recognizes the textual span but wrongly assign the labels to them since some categories provide limited semantic features in the support set to be used in the training stage.

6 Conclusion

In this paper, we address the problem of recognizing the unseen entity categories in the genomic cancer literature. We exploit the few shot learning paradigm in this task and propose a transited pro-

[2] https://pytorch.org/docs/stable/distributions.html

totype NER framework to generate the transition scores for CRF models. Meanwhile, since the training samples are limited in the support set, which results in data fluctuation, we adopt Gaussian distribution as our transition function to smoothen the randomness caused by a few samples. Finally, experimental results on the three cancer genomic datasets prove the effectiveness of our proposed method.

References

Johannes Birgmeier, Cole A Deisseroth, Laura E Hayward, Luisa MT Galhardo, Andrew P Tierno, Karthik A Jagadeesh, Peter D Stenson, David N Cooper, Jonathan A Bernstein, Maximilian Haeussler, et al. 2020. Avada: toward automated pathogenic variant evidence retrieval directly from the full-text literature. *Genetics in Medicine*, 22(2):362–370.

Jari Björne and Tapio Salakoski. 2018. Biomedical event extraction using convolutional neural networks and dependency parsing. In *Proceedings of the BioNLP 2018 workshop*, pages 98–108.

Jiarun Cao, Elke M van Veen, Niels Peek, Andrew G Renehan, and Sophia Ananiadou. 2021. Epicure: Ensemble pretrained models for extracting cancer mutations from literature. In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 461–467. IEEE.

J Gregory Caporaso, William A Baumgartner Jr, David A Randolph, K Bretonnel Cohen, and Lawrence Hunter. 2007. Mutationfinder: a high-performance system for extracting point mutation mentions from text. *Bioinformatics*, 23(14):1862–1865.

Juan Miguel Cejuela, Aleksandar Bojchevski, Carsten Uhlig, Rustem Bekmukhametov, Sanjeev Kumar Karn, Shpend Mahmuti, Ashish Baghudana, Ankit Dubey, Venkata P Satagopam, and Burkhard Rost. 2017. nala: text mining natural language mutation mentions. *Bioinformatics*, 33(12):1852–1858.

Xin Cong, Shiyao Cui, Bowen Yu, Tingwen Liu, Yubin Wang, and Bin Wang. 2020. Few-shot event detection with prototypical amortized conditional random field. *arXiv preprint arXiv:2012.02353*.

Zhenjin Dai, Xutao Wang, Pin Ni, Yuming Li, Gangmin Li, and Xuming Bai. 2019. Named entity recognition using bert bilstm crf for chinese electronic health records. In *2019 12th international congress on image and signal processing, biomedical engineering and informatics (cisp-bmei)*, pages 1–5. IEEE.

Sarkar Snigdha Sarathi Das, Arzoo Katiyar, Rebecca J Passonneau, and Rui Zhang. 2021. Container: Few-shot named entity recognition via contrastive learning. *arXiv preprint arXiv:2109.07589*.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Hai-Tao Zheng, and Zhiyuan Liu. 2021. Few-nerd: A few-shot named entity recognition dataset. *arXiv preprint arXiv:2105.07464*.

Emily Doughty, Attila Kertesz-Farkas, Olivier Bodenreider, Gary Thompson, Asa Adadey, Thomas Peterson, and Maricel G Kann. 2011. Toward an automatic method for extracting cancer-and other disease-related point mutations from the biomedical literature. *Bioinformatics*, 27(3):408–415.

G David Forney. 1973. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278.

Alexander Fritzler, Varvara Logacheva, and Maksim Kretov. 2019. Few-shot classification in named entity recognition task. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 993–1000.

Yutai Hou, Wanxiang Che, Yongkui Lai, Zhihan Zhou, Yijia Liu, Han Liu, and Ting Liu. 2020. Few-shot slot tagging with collapsed dependency transfer and label-enhanced task-adaptive projection network. *arXiv preprint arXiv:2006.05702*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Viet Dac Lai, Franck Dernoncourt, and Thien Huu Nguyen. 2020. Extensively matching for few-shot learning event detection. *arXiv preprint arXiv:2006.10093*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Kyubum Lee, Sunwon Lee, Sungjoon Park, Sunkyu Kim, Suhkyung Kim, Kwanghun Choi, Aik Choon Tan, and Jaewoo Kang. 2016. Bronco: Biomedical entity relation oncology corpus for extracting gene-variant-disease-drug relations. *Database*, 2016.

Jing Li, Billy Chiu, Shanshan Feng, and Hao Wang. 2020. Few-shot named entity recognition via meta-learning. *IEEE Transactions on Knowledge and Data Engineering*.

Ling Luo, Zhihao Yang, Pei Yang, Yin Zhang, Lei Wang, Hongfei Lin, and Jian Wang. 2018. An attention-based bilstm-crf approach to document-level chemical named entity recognition. *Bioinformatics*, 34(8):1381–1388.

Tingting Ma, Huiqiang Jiang, Qianhui Wu, Tiejun Zhao, and Chin-Yew Lin. 2022. Decomposed meta-learning for few-shot named entity recognition. In

Findings of the Association for Computational Linguistics: ACL 2022, pages 1584–1596, Dublin, Ireland. Association for Computational Linguistics.

ASM Ashique Mahmood, Tsung-Jung Wu, Raja Mazumder, and K Vijay-Shanker. 2016. Dimex: a text mining system for mutation-disease association extraction. *PloS one*, 11(4):e0152725.

Chris McCormick. 2016. Word2vec tutorial-the skip-gram model. *Apr-2016.[Online]. Available: http://mccormickml. com/2016/04/19/word2vec-tutorial-the-skip-gram-model*.

Arzucan Özgür, Thuy Vu, Güneş Erkan, and Dragomir R Radev. 2008. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. *Bioinformatics*, 24(13):i277–i285.

Sunil Kumar Sahu, Fenia Christopoulou, Makoto Miwa, and Sophia Ananiadou. 2019. Inter-sentence relation extraction with document-level graph convolutional neural network. *arXiv preprint arXiv:1906.04684*.

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the international joint workshop on natural language processing in biomedicine and its applications (NLPBA/BioNLP)*, pages 107–110.

Yuqi Si and Kirk Roberts. 2018. A frame-based nlp system for cancer-related information extraction. In *AMIA annual symposium proceedings*, volume 2018, page 1524. American Medical Informatics Association.

Jake Snell, Kevin Swersky, and Richard Zemel. 2017. Prototypical networks for few-shot learning. *Advances in neural information processing systems*, 30.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2020. Linear-time wordpiece tokenization. *arXiv e-prints*, pages arXiv–2012.

Flood Sung, Yongxin Yang, Li Zhang, Tao Xiang, Philip HS Torr, and Timothy M Hospedales. 2018. Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1199–1208.

Hai-Long Trieu, Thy Thy Tran, Khoa NA Duong, Anh Nguyen, Makoto Miwa, and Sophia Ananiadou. 2020. Deepeventmine: end-to-end neural nested event extraction from biomedical texts. *Bioinformatics*, 36(19):4910–4917.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. 2016. Matching networks for one shot learning. *Advances in neural information processing systems*, 29.

Hanna M Wallach. 2004. Conditional random fields: An introduction. *Technical Reports (CIS)*, page 22.

Rui Wang, Tong Yu, Handong Zhao, Sungchul Kim, Subrata Mitra, Ruiyi Zhang, and Ricardo Henao. 2022. Few-shot class-incremental learning for named entity recognition. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 571–582.

Yaqing Wang, Subhabrata Mukherjee, Haoda Chu, Yuancheng Tu, Ming Wu, Jing Gao, and Ahmed Hassan Awadallah. 2021. Meta self-training for few-shot neural sequence labeling. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1737–1747.

Chih-Hsuan Wei, Bethany R Harris, Hung-Yu Kao, and Zhiyong Lu. 2013. tmvar: a text mining approach for extracting sequence variants in biomedical literature. *Bioinformatics*, 29(11):1433–1439.

Chih-Hsuan Wei, Hung-Yu Kao, and Zhiyong Lu. 2015. Gnormplus: an integrative approach for tagging genes, gene families, and protein domains. *BioMed research international*, 2015.

Chih-Hsuan Wei, Lon Phan, Juliana Feltz, Rama Maiti, Tim Hefferon, and Zhiyong Lu. 2018. tmvar 2.0: integrating genomic variant information from literature with dbsnp and clinvar for precision medicine. *Bioinformatics*, 34(1):80–87.

Yi Yang and Arzoo Katiyar. 2020. Simple and effective few-shot named entity recognition with structured nearest neighbor learning. *arXiv preprint arXiv:2010.02405*.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. Docred: A large-scale document-level relation extraction dataset. *arXiv preprint arXiv:1906.06127*.

Deyu Zhou, Dayou Zhong, and Yulan He. 2014. Biomedical relation extraction: from binary to complex. *Computational and mathematical methods in medicine*, 2014.

7 Appendix

Table 4 and Table 5 are shown in the next page.

		1Way-6Shot-10Epoch			*2Way-6Shot-10Epoch*			*3Way-6Shot-10Epoch*			*1Way-6Shot-100Epoch*			*2Way-6Shot-100Epoch*			*3Way-6Shot-100Epoch*		
	Models	Prec.	Rec.	F1 score	Prec.	Rec.	F1 score	Prec.	Rec.	F1 score	Prec.	Rec.	F1 score	Prec.	Rec.	F1 score	Prec.	Rec.	F1 score
Vanilla NER models	BiLSTM+CRF	5.0%	80.0%	9.4%	8.0%	7.0%	14.4%	0.0%	0.0%	0.0%	5.0%	90.0%	9.5%	2.0%	**90.0%**	3.9%	0.0%	0.0%	0.0%
	BERT+CRF	3.00%	**90.00%**	5.80%	3.00%	**90.00%**	5.8%	0.0%	0.0%	0.0%	4.0%	67.0%	7.5%	2.0%	**90.0%**	3.9%	0.0%	0.0%	0.0%
FSL models in general	Matching Network	5.12%	50.00%	9.29%	6.88%	75.00%	12.61%	5.12%	50.0%	9.29%	7.13%	**96.00%**	13.28%	4.21%	50.50%	7.77%	3.45%	**35.67%**	6.29%
	Relation Model	12.41%	25.64%	16.72%	8.23%	24.61%	12.33%	7.61%	21.11%	11.19%	9.18%	10.00%	9.57%	5.21%	28.00%	8.79%	3.45%	16.67%	5.72%
	Proto Network	17.24%	50.00%	25.64%	12.31%	40.00%	18.82%	17.24%	50.00%	25.64%	7.04%	24.00%	10.88%	5.70%	26.50%	9.38%	4.28%	18.67%	6.97%
	Proto+Dot	22.80%	48.00%	30.92%	19.70 %	40.00%	26.40%	16.20%	36.70%	24.48%	7.12%	26.40%	11.22%	7.12%	24.00%	10.98%	6.12%	16.67%	8.95%
	Proto+Att	23.00%	46.00%	30.67%	20.10%	42.00%	27.19%	17.24%	36.00%	23.30%	9.18%	28.00%	13.83%	7.31%	24.12%	11.22%	5.82%	16.67%	8.62%
CRF-based models	CONTAINER (Das et al., 2021)	27.50%	52.34%	35.48%	-	-	-	-	-	-	**32.26%**	45.45%	37.74%	-	-	-	-	-	-
	FEW-NERD (Ding et al., 2021)	24.90%	68.39%	36.50%	19.63%	68.54%	30.52%	15.70%	**61.15%**	24.98%	26.57%	82.45%	40.19%	**24.52%**	20.19%	22.15%	22.11%	6.66%	10.24%
	DML (Ma et al., 2022)	25.31%	51.67%	33.98%	**38.12%**	34.12%	36.01%	14.89%	50.23%	22.97%	29.74%	69.23%	41.61%	19.35%	26.78%	22.47%	**22.43%**	6.12%	9.62%
	our model	**31.25%**	50.00%	**38.46%**	30.30%	50.00%	**37.74%**	**23.08%**	30.00%	**26.09%**	32.20%	66.00%	**43.28%**	17.14%	36.50%	**23.32%**	8.51%	20.00%	**11.94%**

Table 4: Precision, recall and F1 scores of different models on TmVar dataset.
Bold marks the highest figure, underline marks the second-highest figure.

		1Way-5Shot-5Epoch			*1Way-5Shot-10Epoch*			*1Way-1Shot-5Epoch*			*2Way-1Shot-5Epoch*			*2Way-1Shot-10Epoch*			*2Way-5Shot-5Epoch*		
		Prec.	Rec.	F1 score	Prec.	Rec.	F1 score	Prec.	Rec.	F1 score	Prec.	Rec.	F1 score	Prec.	Rec.	F1 score	Prec.	Rec.	F1 score
Vanilla NER models	BiLSTM+CRF	10.00%	**100.00%**	18.18%	1.00%	90.00%	19.78%	4.00%	100.00%	7.69%	0.00%	0.00%	0.00%	2.00%	**100.00%**	3.92%	5.00%	80.00%	04.12%
	BERT+CRF	4.00%	**100.00%**	7.32%	1.00%	90.00%	0.20%	4.00%	**100.00%**	7.32%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	2.00%	78.00%	4.09%
FSL models in general	Matching Network	8.47%	**100.00%**	15.63%	4.93%	**100.00%**	9.39%	6.67%	**100.00%**	12.50%	1.01%	40.00%	1.96%	2.84%	90.00%	5.51%	3.13%	**90.00%**	6.04%
	Relation Model	9.12%	90.00%	16.56%	4.35%	90.00%	8.30%	6.67%	100.00%	12.51%	1.41%	40.00%	2.72%	3.12%	95.00%	6.04%	4.41%	80.00%	8.36%
	Proto Network	11.11%	20.00%	14.29%	5.33%	40.00%	9.41%	5.15%	100.00%	9.80%	3.05%	90.00%	5.90%	2.75%	95.00%	5.34%	2.99%	60.00%	5.69%
	Proto+Dot	16.20%	25.00%	19.66%	6.89%	36.00%	11.58%	8.12%	90.00%	14.90%	7.21%	90.00%	13.35%	2.75%	95.00%	5.34%	12.08%	60.00%	20.11%
	Proto+Att	18.12%	20.00%	19.01%	5.82%	40.00%	10.16%	7.19%	**100.00%**	13.42%	6.28%	**100.00%**	11.82%	3.56%	96.00%	6.87%	13.01%	50.00%	**20.65%**
CRF-based models	CONTAINER (Das et al., 2021)	24.84%	32.77%	28.26%	18.56%	55.31%	27.79%	-	-	-	17.65%	13.22%	15.12%	21.04%	23.41%	22.16%	-	-	-
	FEW-NERD (Ding et al., 2021)	19.27%	51.30%	28.02%	16.16%	54.71%	24.95%	21.45%	37.45%	27.28%	12.75%	20.27%	15.65%	14.80%	54.72%	23.21%	10.25%	36.24%	13.86%
	DML (Ma et al., 2022)	25.80%	30.12%	27.79%	20.12%	48.94%	**28.52%**	33.80%	16.67%	22.33%	**6.12%**	66.67%	11.21%	17.43%	33.24%	22.87%	7.24%	34.12%	11.95%
	Our model	**50.00%**	20.00%	**28.50%**	**50.00%**	20.00%	**28.50%**	**50.00%**	20.00%	**28.50%**	**50.00%**	10.00%	**16.67%**	**33.34%**	20.17%	**25.00%**	**50.00%**	10.00%	16.70%

Table 5: Precision, recall and F1 scores of different models on BRONCO & EMU dataset.
Bold marks the highest figure, underline marks the second-highest figure.