**Cover Letter**

30th August, 2020

Dear Editors and Reviewers,

We would like to express sincere gratitude and appreciation toward the editors and peer reviewers who spent time reviewing our manuscript and provided critical comments to our paper. We have made significant changes to improve the quality of our paper to a level we believe will be satisfactory for the peer reviewer. All reviewers’ comments are responded, and revisions are made accordingly. For clarity, we have color-coded texts as follows:

* The reviewer’s comments in green;
* Our reply in **black;**
* Revised sentences in the manuscript in blue**;**

All changes made in the revised version of the manuscript are outlined in red.

Thank you for your consideration for this manuscript.

Sincerely,

Authors

**Response to the Reviewers’ Comments**

**Response to Reviewer 1**

**Comment 1:**

Using the traditional method, similar to clustering, on the same dataset can only get less than 20% accuracy. This confirms that our proposed method is better than traditional methods. -> The comparative work seems inadequate. Target related to clustering is not clear.

**Response 1:**

Thank you very much for your comment, we apologize that our aim lost its captivating style. We do agree that the description of the FDC method is not clear, and the presentation is not clear enough. In fact, we added a comparison method in section 2, which describes the principle of the FDC method. And we revised this sentence:

Since there are few related work that can use our data set, we almost explored it ourselves. There are very few methods that can be used on our data set, but we still found a traditional method called FDC analysis. We reproduced this method and used it on our data set. The final accuracy rate achieved was only 26.41. %. This accuracy is far less than the accuracy that DPNet can achieve.

Because the characteristics of our data are only diagnosis and prescription, they are completely discrete and without distance, and we can only find one type of FDC after consulting the data to process such data. So in such a situation, our method is novel, and related work is rare. Compared with FDC, the results are also quite excellent, so it can be considered that the effect shown by DPNet is good.

**Comment 2:**

References are too old: 2008 and 2009.

**Response 2:**

Thank you very much for your comment. We updated most of the old references, but one reference had to be left: cite 5. Because related work is very rare, and we only use the two-dimensional features of diagnosis and prescription, other methods are not applicable to our data . And the method in this reference is used by us for a comparison, so this reference needs to be kept. We update the reference:

We adopt five key delay propagation mechanisms focused on how the delay of an example train, train 1 travelling from station A to Station B propagates to some train 2, with the exception of delay propagation mechanism 1 which concerns train 1 exclusively. The pseudo code for the allocation of delay propagation minutes to train journeys for mechanisms 2 to 5 are detailed in Appendix A1 to A4..

**Comment 3:**

In Table 1, terms are not described: RN1, AKC196, BKE316 ...

**Response 3:**

Thank you very much for your comment. We apologize for the confusion and we understand that our claim of the importance of a tabular data format is irrelevant. We added the explanation in the bottom of the Table 1.

**Comment 4:**

L 134-141: Similar results, but one is that cross-validation verifies effect, and the other is that cross-validation shows little impact. The explanation is too short.

**Response 4:**

Thank you very much for your insight. What we want to express is that cross-validation has little effect on our experiment, and its influence on the experimental results is about 0.1%. We have written the description in more detail.

The main idea of ​​cross-validation is to divide the data into very small parts. In each cycle, most of them are selected as the training set, and the rest is the validation set. The model may show different results in different schemes. We used cross-validation to carry out the experiment, but it had little effect on the results, only affecting about 0.1% of ADAccurary.

**Comment 5:**

Descriptions for Table 4 and Table 5 are not enough.

Descriptions for Figure 4 and Figure 5 are not enough.

**Response 5:**

Thank you very much for your concern. We provide more details.

Table 4: FL indicates that DPNet uses focal loss and removes the residual channel, CE indicates that DPNet use the residual channel and cross-entropy loss instead focal loss, and the NONE metric uses the cross-entropy loss and does not use the residual channel. The experimental results clearly show that the model using focal loss has the best results, while the model using residual channel has a slight improvement in comparison.

Coresponding section 3.4: Our model uses focal loss to reduce the impact of the imbalance of positive and negative samples, and residual channel to reduce the deep network problem. In order to verify the effectiveness of these two auxiliary methods, we conducted a set of comparative experiments.Three more experiments were done for verifying that focal loss and residual channel are beneficial. They are named a.FL, b.CE, and c.NONE, respectively. They respectively indicate that DPNet removes the residual channel and uses only the focal loss, removes the focal loss and uses only the residual channel, and removes both.

The recall rate of the three experiments is a, b, and c in descending order. However, the precision rate of the three experiments is c, b, and a in descending order. Experimental results show that focal loss can greatly improve the model. Whereas the residual channel does not have much effect on the result. TABLE 4 presents the results. In sub-sub-section 3.3.2, the model uses both focal loss and residual channel. It can be seen that the result is better than the experiment FE. The experimental results show that focal loss ans residual channel both can improve the model, but focal loss has the greatest impact on our model, while residual channel has a slight impact.

Table 5: For the Comparative Experiments, we did an FDC (traditional method) experiment, and a reduced experiment. The specific method of FDC is described in section 2.3, I hope you can move to where to read how FDC does it. The reduced experiment is to reduce the types of diseases and medications in the data set, reducing the original 1,000 types to 100 types. The purpose of this is to verify that our model can learn a better result. To further explain, the original 1000 types may be too many, and there will be many many-to-many relationships, which increases the difficulty of model training. For example, medications a, b and corresponding diseases d, e, when the model training a to d increases the probability of d and reduces the probability of e. Then the model training b to e increases the probability of e and reduces the probability of d. The reduced experiment reduces this many-to-many relationship, and the experimental results clearly show that the effect of the model can be improved.

It should also be noted that the data set used in the FDC experiment is the same as the data set used in the reduced experiment. The FDC method is not only inferior to DPNet in the reduced data set, but the experimental result of only 100-kinds-dataset is not as good as the experimental results of 1000-kinds-dataset.

This is our revised version：

To verify this work is feasible and reliable, we did several additional experiments.

In the experiments of this section, we used a reduced data set. The data set only contains 100 diseases and corresponding medications.

We designed a reduced set experiment, which is also training DPNet, but the data set is replaced with a reduced data set. DPNet uses a normal data set (its diseases and medications are 1000 kinds) can reach about 50\% of the recall, this value is not satisfactory to us, and can not be very convincing to express the model is effective. Therefore, we designed this experiment with the purpose of observing whether DPNet can perform better on a small data set, and whether it can demonstrate reliable learning ability, so as to verify that DPNet can indeed learn effectively.

In order to verify that our model is reliable, we reproduced the FDC analysis method to compare with DPNet. FDC analysis is a traditional prescription anomaly detection method we found. The specific description of the method is in section 2.3.

The experimental results are shown in Table 5.

The reduced experiment showed better results than the normal experiment (Diagnosis-to-Prescription), and it greatly improved recall and ADAccurary. This also shows that DPNet can show good results on small-scale data sets, and it also proves that DPNet can effectively learn the relationship between diagnosis and prescription.

DPNet's ADAccurary and FDC's accuracy can be compared with each other. And our model shows better results. The FDC method is not only inferior to DPNet in the reduced data set, but the experimental result of only 100-kinds-data set is not as good as the experimental results of 1000-kinds-dataset.

For Figure 4.(now change to Figure 5):

The left side of the figure is the model we currently use, and the right side is the improvement that may be improved in the future. For the sequence model on the right. Since there may also be associations between medications, this sequence model can increase the association of medications. For example, the use of α medication alone may be an abnormal situation, but the combination with α medication β becomes a normal situation. And the combination of different medications can affect each other's probability, which is what the sequence model can learn in theory. However, since this sequence has a sequence, it may need to be determined manually, which is currently impossible for us.

For Figure 5.(now change to Appendix.A.1):

System Structure diagram. DPNet is deployed in the monitoring system as the main detection method. \hl{RDBMS is a relational database management system, which mainly stores user information and related information. Relational databases such as MySQL are mainly used. BigDataCentre processes and stores large amounts of data. It uses NOSQL databases, such as mangoDB, Hive, etc., big data processing tools can use Hadoop, Hive, etc. task scheduler, which aims to achieve reliable message transmission and load balancing. Communication encryption ensures that data is not leaked and protects user privacy. MQ is a message queue. It is our main communication method. It can be imagined as a queue in which messages are queued. Different system services put messages in and out messages.

**Comment 6:**

A mixed notation such as Figure 4 or Fig. 5. The author's guide should be checked.

**Response 6:**

We have checked and unified the standard notaion.

**Comment 7:**

Figure 5 shows the system structure, but there are no explanations for each component: H5, Static resources, indicator, Diary, and so forth.

Readers may be not experts in this field.

**Response 7:**

PC means a personal computer, APP means mobile terminal application, H5 means web application made by HTML5 standard which is a cross-platform application. Static Resources represent resources that will not change, such as pictures and text which will not change. Nginx is a reverse proxy server, which can be simply understood as a proxy server that helps manage distributed servers.

**Comment 8:**

It needs to justify the novelty of the proposed scheme, through full explanations and result interpretation. Overall, it's hard to find out the contribution of research because of the abstraction and simplified expression.

**Response 8:**

Thank you very much for your comment. We apologized for causing you confusing. We have revised the article to describe the work of this article as clearly as possible.

Our job is to use only diagnosis and prescriptions to detect abnormalities in prescriptions while protecting user privacy. Our contribution is to achieve anomaly detection under limited conditions, which other researchers have not done. And the accuracy rate obtained can far exceed our traditional method of comparison.

**Response to Reviewer 2**

**Comment 1:**

The sub-section 3 2.1. Overview needs the dataset description should do before it, in order to understand the method. So, I suggest to have a sub-section before, describing the dataset.

**Response 1:**

Thank you very much for your comment, we apologize that our aim lost its captivating style. We do agree that the description of the FDC method is not clear, and the presentation is not clear enough. In fact, we added a comparison method in section 2, which describes the principle of the FDC method. And we revised this sentence:

**Comment 2:**

The model description should be re-structured. I suggest to firstly present the whole model, then present the layers.

**Response 2:**

Thank you very much for your comments, what we described may indeed be unclear. We have modified this sequence to make it more clear to show the model structure.

**Comment 3:**

About the model: why does mean having a PRESCRIPTION layer at the end? Authors should explain it better.

**Response 3:**

Thank you very much for your comment. This prescription is the output of the model, which is actually the probability of 1000 medications. We use the label( the corresponding presscription) and this probability matrix to do back propagation or judge whether it is abnormal.

In fact, DPNet derives the probability of all medications (1000 kinds), and then selects the corresponding probability according to the label of the original data (what medications are used), that is, the probability of obtaining the entire prescription.

**Comment 4:**

How the data as labeled? Manually?

**Response 4:**

Thank you very much for your comment, we apologize for not describing this problem clearly. We do agree that the description of How the data as labeled is not clear. We have added descriptions in the abstract and introduction and results of the article, and our data is unlabeled. Because the amount of data is large and there are few abnormal data, we can treat all data as normal data, so that the relationship between diagnoses and prescriptions that model learning is correct in terms of probability.

**Comment 5:**

Sub-section 3.2. Pre-processing is in the middle of results. Put it close to the dataset description.

**Response 5:**

Thank you for your suggest. We changed the order.

**Comment 6:**

Sub-sections 3.3.1. Diagnosis-to-Prescription and 3.3.2. Prescription-to-Diagnosis do not make any sense. Authors should explain what is the reason of these experiments and why they were delineated.

**Response 6:**

Thank you very much for your comment, we apologize that our description is not clear. We have given the specific reasons and the rationale for why this is done.

Under normal circumstances, we get the probability of a prescription through diagnosis. We denote this experiment as Diagnosis-to-Prescription. We found that multiple different prescriptions may correspond to the same diagnosis. This is not good for the training of the model. If we can reduce the many-to-many relationship, then the model can be optimized. Thus, we camp up an idea of swapping input and output to reduce this many-to-many relationship. We denote the swapping experiment as Prescription-to-Diagnosis.

**Comment 7:**

The description of the results presented in sub-section 3.4. Verification Experiments is very hard to understand. The same for sub-section 3.5. Comparative Experiments.

**Response 7:**

Thank you for your suggest. We reorganized the language and described the two sets of experiments in more detail.

For the Verification Experiments, Our main purpose is to verify that focal loss and residual channel have improved our model.

Our model uses focal loss to reduce the impact of the imbalance of positive and negative samples, and residual channel to reduce the deep network problem. In order to verify the effectiveness of these two auxiliary methods, we conducted a set of comparative experiments.

For the Comparative Experiments, we did an FDC (traditional method) experiment, and a reduced experiment. The specific method of FDC is described in section 2.3, I hope you can move to where to read how FDC does it. The reduced experiment is to reduce the types of diseases and medications in the data set, reducing the original 1,000 types to 100 types. The purpose of this is to verify that our model can learn a better result. To further explain, the original 1000 types may be too many, and there will be many many-to-many relationships, which increases the difficulty of model training. For example, medications a, b and corresponding diseases d, e, when the model training a to d increases the probability of d and reduces the probability of e. Then the model training b to e increases the probability of e and reduces the probability of d. The reduced experiment reduces this many-to-many relationship, and the experimental results clearly show that the effect of the model can be improved.

It should also be noted that the data set used in the FDC experiment is the same as the data set used in the reduced experiment. The FDC method is not only inferior to DPNet in the reduced data set, but the experimental result of only 100-kinds-dataset is not as good as the experimental results of 1000-kinds-dataset.

This is our revised version：

To verify this work is feasible and reliable, we did several additional experiments.

In the experiments of this section, we used a reduced data set. The data set only contains 100 diseases and corresponding medications.

We designed a reduced set experiment, which is also training DPNet, but the data set is replaced with a reduced data set. DPNet uses a normal data set (its diseases and medications are 1000 kinds) can reach about 50\% of the recall, this value is not satisfactory to us, and can not be very convincing to express the model is effective. Therefore, we designed this experiment with the purpose of observing whether DPNet can perform better on a small data set, and whether it can demonstrate reliable learning ability, so as to verify that DPNet can indeed learn effectively.

In order to verify that our model is reliable, we reproduced the FDC analysis method to compare with DPNet. FDC analysis is a traditional prescription anomaly detection method we found. The specific description of the method is in section 2.3.

The experimental results are shown in Table 5.

The reduced experiment showed better results than the normal experiment (Diagnosis-to-Prescription), and it greatly improved recall and ADAccurary. This also shows that DPNet can show good results on small-scale data sets, and it also proves that DPNet can effectively learn the relationship between diagnosis and prescription.

DPNet's ADAccurary and FDC's accuracy can be compared with each other. And our model shows better results. The FDC method is not only inferior to DPNet in the reduced data set, but the experimental result of only 100-kinds-data set is not as good as the experimental results of 1000-kinds-dataset.

**Comment 8:**

There are some results in the discussion that were not presented before: "In most experiments, loss convergent 163 rapidly in the beginning, but in the middle and later stages its loss fluctuates between 10 and 30."

**Response 8:**

Thank you for pointing out our mistake, this is indeed our negligence. We have added the loss curve of swapping experiments and added instructions.

**Comment 9:**

The section 5. Application is not enough described and it seems not necessary.

**Response 9:**

Thank you for your suggestion, we agree with your point of view, this Application is indeed a bit abrupt. But in fact, we are cooperating with enterprises, so this project should eventually be applied in practice. But it is a bit reluctant to appear in the main body, so we decided to move it to the appendix.

In addition, we describe this content in more detail.

**Comment 9:**

At the end of the paper, authors propose a FC layer based model, because they say that their proposal is still a bit inadequate. The main question is: why not present this new and probably better model instead this one?

**Response 9:**

Thanks for your suggestion. We should apologize for this oversight. We should have added this explanation originally, but considering that it may occupy a lot of space, there is no. This makes readers suffer some confusion, which is what we don't want to see. So, we explain why the new method is not used.

This sequential model can theoretically improve the accuracy of the model, but it has some limitations. This model is sequence-related, which means that we must determine a set of sequences. Our original intention is that the use of medications may be a combination. It may be abnormal to use a certain medication alone, but it is normal to use it in combination with other medications. But the problem is that we don't know the order of this relationship, which may require manual labels. The content we added is as follows:

But this model is based on sequences, and different sequences will have different results. Therefore, we may need manual labels, which is not advisable for us at present, and this method here is just to propose a possible idea.

**Response to Reviewer 3**

**Comment 1:**

a) doing a more thorough comparison with other methods of detecting prescription abnormality; one comparative experiment (presented in section 3.5) is really not enough, especially since it is not clear what this experiment is, nor why a "reduced dataset" was used. Why is ADAaccuracy used, if this is not a standard method in the field? Much more must be done to show that this method outperforms the other methods, as claimed.

**Response 1:**

Thank you very much for your comment, we apologize for the unclear points in the article. Regarding doing a more comparison, unfortunately, we tried our best to find only one method that can be used on our data set. We did not find it more because our data features are too few, and other methods require Additional information such as amount, frequency of visits, etc. So many methods are not applicable. We describe the comparison method (FDC) in more detail, in section 2.3. However, we found another one that is very similar to our work, but unfortunately their data set has an extra column(amount) of data, which we cannot do. We also put this into the comparison.

Another method is the abnormal medical prescription detection method based on the extended topic model [6]. The data set in this method is very close to our data, but we cannot use this method because this method requires the use of amount information, and our faced challenge is to use only diagnosis and prescription.

We found the research of another article. They are based on topic-based prescription anomaly detection [6], but they use one more dimension of feature amount than ours. Although we can't reproduce their method on our data set, the similar work can still be compared. The table also shows the accuracy and recall rate of the other party. And from the result, our method is better.

Why use a "reduced dataset". We designed this experiment with the purpose of observing whether DPNet can perform better on a small data set, and whether it can demonstrate reliable learning ability, so as to verify that DPNet can indeed learn effectively. And we added this sentence:

We designed a reduced set experiment, which is also training DPNet, but the data set is replaced with a reduced data set. DPNet uses a normal data set (its diseases and medications are 1000 kinds) can reach about 50% of the recall, this value is not satisfactory to us, and can not be very convincing to express the model is effective. Therefore, we designed this experiment with the purpose of observing whether DPNet can perform better on a small data set, and whether it can demonstrate reliable learning ability, so as to verify that DPNet can indeed learn effectively.

Why is ADAaccuracy used, if this is not a standard method in the field? Because our model requires an anomaly indicator, and the existing anomaly indicator is not applicable, our ADAaccuracy is not completely created by ourselves, but based on recall. The reason why we only use recall is: due to the many-to-many relationship, precision does not have much meaning. There are many possible medications for the same disease, and these are all correct, so precision loses the reliability of judging abnormalities. We use ADAaccuracy to filter out suspicious data. ADAaccuracy is also for the convenience of referencing this standard later (recall>0.5), reducing unnecessary explanations. And we revised this sentence:

ADAccurary means the proportion of data whose recall is greater than 0.5, that is, the proportion of normal data. Because we treat all data as normal data, ADAccurary can also be used to express the accuracy of our model.

**Comment 2:**

b) The authors should provide evidence (statements with references to relevant papers are sufficient) that prescription abnormalities are an important source of fraud.

**Response 2:**

Thank you very much for your comment, we added some new content and revised some content to explain prescription abnormalities are an important source of fraud.

Medical insurance fraud is considered from the perspective of human behavior, so it is possible that human behavior is fraudulent. From the data itself, it may be an abnormal prescription. Abnormal prescriptions are a source of health insurance fraud. A study shows that the UK spends 256 million pounds on prescription fraud every year[26]. Prescription fraud and abuse have been a pressing issue in the USA, resulting in large financial losses and adverse effects on human health [27]. Fraud in the form of abnormal prescriptions is not only a personal act, but also deceives every country with a medical insurance system.

Now we have a demand, because we need to protect user privacy, so we have to judge whether the prescription is abnormal under the condition of only diagnosis and prescription.