International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

Reducing Features Within an Extensive Network of Medical Subject Headings Metadata to Enhance Deep Predictions

Zolzaya Dashdorj[1,2]*, Zoljargal Jargalsaikhan[1], Stanislav Grigorev[2], Andrey Trufanov[2], Erdenebaatar Altangerel[1]
[1]Mongolian University of Science and Technology, Ulaanbaatar, Mongolia
[2]Irkutsk National Research Technical University; Irkutsk, Russia
[1]zolzaya@must.edu.mn*, j.zoljargal@must.edu.mn, erka@must.edu.mn
[2]svg@istu.edu, troufan@istu.edu

Abstract—Feature reduction in a large amount of Medical Subject Headings (MeSH) metadata for deep prediction involves selecting a subset of relevant features to improve the efficiency and effectiveness of deep learning models. MeSH is a controlled vocabulary thesaurus used for indexing articles in the life sciences and biomedical fields. We analyze a disease-symptom network exploiting MeSH metadata and configure a deep model for disease prediction based on symptoms. Dimension reduction techniques have yielded positive results in optimizing a large amount of Medical Subject Headings (MeSH) metadata for deep prediction with good accuracy. Therefore, our result highlights that decrease in the severity or degree of symptoms associated with a disease correlates with an improvement in the accuracy of disease prediction. This finding may have important implications for disease prediction models in healthcare: interpretation of results, clinical significance, practical implications.

Keywords-health-care data analytics; bio-computing; deep learning; disease network; optimization

I. INTRODUCTION

The swift progress of machine learning and data analytics has opened the door to inventive strategies in disease prediction. One promising approach involves utilizing a network-based method that exploits symptom data to predict and comprehend disease outcomes [1,9,10]. This method integrates various data sources and establishes connections between symptoms and diseases, providing a thorough insight into the intricate interactions influencing human health. Numerous applications have emerged, such as healthcare chatbots and disease diagnosis using CT and MRI images. The integration of multiple data sources is a hallmark of cutting-edge disease prediction models, encompassing electronic health records (EHRs), genomic data, wearable devices, social media, and environmental factors. The amalgamation of these diverse data sets aims to capture a holistic perspective of an individual's health status, ultimately enhancing the precision of predictions. The progression of these methodologies is significantly propelled by the utilization of machine learning and deep learning techniques. Cutting-edge machine learning algorithms [2-8,11-17], including decision trees, random forests, support vector machines, and neural networks, along with sophisticated deep learning models such as convolutional neural networks (CNNs) [2-4,18] and recurrent neural networks (RNNs) [5-6], are leveraged in the realm of disease prediction [10-16]. These models have the capacity to discern intricate patterns and connections within symptom data, facilitating precise predictions [7]. Disease prediction models grounded in network analysis have garnered significant attention. These models encapsulate the intricate relationships and interactions among symptoms, diseases, and biological entities such as genes and proteins [1,7]. Network-based approaches unveil concealed associations, shedding light on disease mechanisms and potential therapeutic targets. Several studies have explored the reduction of features within extensive networks of Medical Subject Headings (MeSH) metadata to enhance deep prediction. The focus has been on leveraging advanced techniques, such as dimensionality reduction and feature selection, to improve the efficiency and accuracy of deep prediction models. These approaches aim to optimize the utilization of MeSH metadata by identifying and retaining the most relevant features for disease prediction tasks. The studies emphasize the importance of selecting an appropriate subset of features to enhance the performance of deep learning models in the context of biomedical and healthcare data. The ultimate goal is to streamline the prediction process and achieve more accurate and efficient outcomes in disease prediction based on MeSH metadata networks. This research endeavor aims to formulate a network-based approach for disease prediction utilizing symptom data. In this investigation, we scrutinize a disease-symptom relation network to comprehend the characteristics of patients' symptoms in terms of occurrence, aiming to refine disease diagnosis and prediction tasks. Our study illustrates the estimation of disease predictability and the correlation between diseases and symptoms through deep learning models. Through the construction of a symptoms-disease network, we can apprehend the interconnectedness between symptoms and diseases, revealing latent patterns and associations that have the potential to enhance diagnostic accuracy and inform treatment decisions.

II. RELATED WORKS

The forefront of disease prediction based on symptoms embraces the use of sophisticated machine learning methods to scrutinize patient data and deliver precise predictions. A noteworthy strategy involves employing deep learning models, including recurrent neural networks (RNNs) and convolutional neural networks (CNNs), which have demonstrated

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

encouraging outcomes in discerning intricate patterns and relationships within symptom data. A research conducted by Choi et al. [5] introduced the Retain model, an interpretable predictive model for healthcare featuring a reverse time attention mechanism. This model assimilates patient symptoms over time to anticipate disease outcomes, facilitating both effective disease prediction and interpretability. The Retain model has showcased its capability to capture longitudinal data and offer insights into the decision-making process. Moreover, the inclusion of electronic health records (EHRs) has emerged as a pivotal element in disease prediction. Scholars like Jensen et al. (2012) have concentrated on extracting valuable information from EHR data to construct robust disease prediction models. These models, utilizing extensive EHR databases, can tap into comprehensive patient details encompassing symptoms, medical history, and clinical measurements, thereby elevating the accuracy of disease prediction. Beyond EHR data, the integration of diverse modalities such as genetic data and medical imaging has garnered attention in disease prediction research. Through the amalgamation of multiple data sources, researchers can amplify the predictive capacity of models and furnish a more holistic comprehension of diseases. In disease prediction based on symptoms, cutting-edge practices encompass the deployment of sophisticated machine learning models, the incorporation of diverse data sources, and the creation of interpretable models. These innovations carry substantial promise for enhancing diagnostic accuracy and facilitating more informed treatment decisions, thereby contributing to improved healthcare outcomes. The research conducted by Zhou et al. [1] delivers a thorough examination of the symptoms-disease network, elucidating the connections between symptoms and diseases. The results underscore the feasibility of leveraging network-based approaches in the realms of disease diagnosis, prediction, and comprehension of disease mechanisms.

III. Methodology

The following 3-stage research was conducted using machine learning.

A. *Analyze the degree of links between diseases and symptoms*

It is crucial to explore hidden links between diseases to understand their characteristic differences. Symptoms could be one of the factors to diagnose the characteristic differences. We analyze common and unusual symptoms in the disease-symptom network based on the degree link between the symptom and diseases.

B. *Estimate the degree of predictability of diseases*

By assessing the distinctiveness of symptoms, our endeavor is to employ a deep learning model. The CNN architecture is structured as a multi-layer network, strategically designed to minimize data processing requirements, as illustrated in Figure 1. Sequentially five layers were designed with the following hyper-parameters. The 1st layer is a one-dimensional convolutional layer of 64 filters, 2 kernels and Relu activation function. We add Dense (16 units), MaxPooling1D and Flatten layers into the model. The output layer contains the number of output classes and 'softmax' activation. The evaluation results of those models are estimated by F1-score, Precision, Recall.

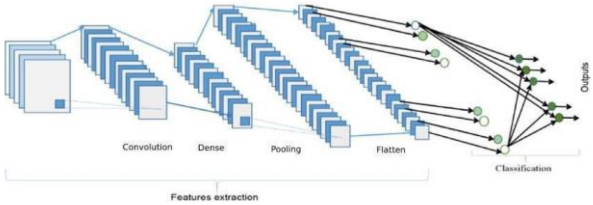

Figure 1. Convolutional Neural Network architecture

C. *Feature reduction of symptoms*

Considering the high-dimensional characteristics of symptom data, we employ feature selection and dimensionality reduction techniques to pinpoint the most informative and pertinent symptoms. Techniques such as genetic algorithms, recursive feature elimination, and L1 regularization are applied in feature selection to eliminate redundant or irrelevant symptoms, thereby improving prediction performance. Diseases often manifest with a broad spectrum of symptoms, exhibiting variations in severity, duration, and combinations. The diversity in symptoms poses a challenge to disease prediction, given that distinct individuals may display different symptom sets for the same disease. Leveraging autoencoder techniques, we reduce the number of symptoms to identify the essential features indicative of symptoms for each disease.

IV. Experimental results

We used PubMed articles with 7,109,429 (about 35.5% of over twenty million records) disease/symptom terms in the MeSH metadata field [1]. This data represents 98.5% of all symptoms and 95.0% of all diseases in the MeSH vocabulary. In this research, we focus on diseases and symptoms. Therefore, the dataset is a total of 4,218 diseases and 322 different symptoms. Therefore, we considered the severity of the symptoms based on the TF/IDF of PubMed occurrences. For instance, the network of diseases and symptoms that sampled around ten thousand tuples is visualized in Figure 2. In total, 4,540 nodes of diseases and symptoms, and 147,977 edges between diseases and symptoms are observed in the network.

Figure 2. Network of diseases and symptoms

The degree distribution of in-degree and out-degree is visualized in plot with log scaling on both the x- and y-axis in Figure 3. The in-degree of symptoms is 1.65 that means a single symptom is linked to 1.65 diseases on average and at

Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 September 2023*

the maximum 3,971 diseases. The out-degree of disease is 16.3 that means a single disease is linked to 16.3 symptoms on average and at the maximum 142 symptoms.

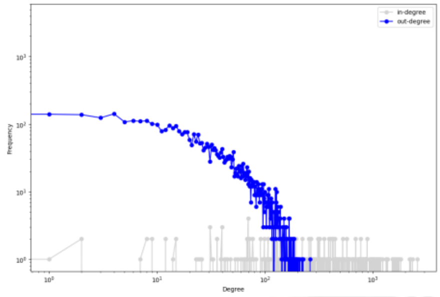

Figure 3. Degree distribution of links between diseases and symptoms

Only 140 diseases are linked to a single symptom. For example, 20 diseases linked to a single symptom are visualized in Figure 4. The nodes in green color of small size represent a disease, and the node in blue color with a larger round describes a symptom. For instance, monkeypox disease and lumpy skin disease are linked to a fever symptom only in MeSH metadata.

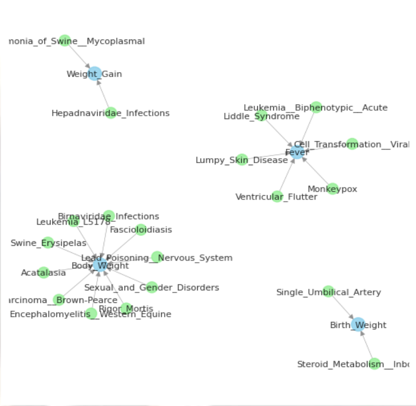

Figure 4. Example diseases linked to a single symptom

Average clustering coefficient is 0.53 that many nodes were connected tightly forming a triangle of clusters. Figure 5 shows the distribution of clustering coefficient of nodes. Our goal is to minimize the similar clusters by selecting important features of symptoms. The distribution of in degree clustering coefficients shows that the characteristic difference of symptoms are high. in terms of link to the diseases. Therefore, out-degree distribution shows that many diseases share the same symptoms.

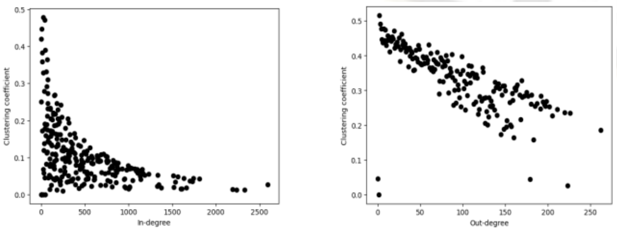

Figure 5. Distribution of clustering coefficients

We evaluate the disease prediction models by F1-score, Precision, and Recall. Logistic regression model performed poorly with the accuracy of 14.0%. Adopting the CNN model, the accuracy of training data is 91.47%. Taking the severity of symptoms into account the CNN model achieved 97.4% of accuracy, 97.3% of precision, and 97.9% of recall in the training set. A large amount of medical Subject Headings (MeSH) metadata requires computational power for prediction. Selecting a subset of relevant features improves the efficiency and effectiveness of deep learning models considering the severeness degree of symptoms. We applied a linear autoencoder to reduce the number of features in symptoms. Figure 6 shows the accuracy of the CNN model given reduced features. Considering at least 84 reduced features in symptoms provides more than 90.18% of accuracy on the CNN model. This reduces the computational cost of deep models by minimizing the symptoms. For instance, at the maximum, 142 symptoms linked to a single disease.

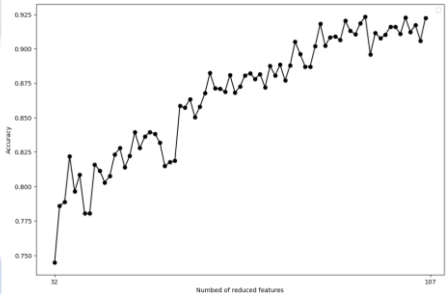

Figure 6. Accuracy of the CNN model in reduced number of features

This trained model would be used to further predict diseases from any form of medical documents such as clinical notes. But the terms should be aligned with the MeSH Disease Terms and MeSH Symptoms Terms. In order to validate our CNN model configuration, we also demonstrated the experiment on clinical data records [17]. Using clinical notes used in [17], a total of 4,920 patient records were obtained in this research. The dataset of 41 types of disease consisting of 135 symptoms was used, and the degree of symptom severity was graded on three levels. Every disease in our dataset is associated with up to 18 symptoms. The dataset were well-balanced. The model performance was evaluated in split data 80/20. Given symptoms, predicting a particular disease adopting the CNN model that we configured is 99.9% of the F1-score. The evaluation results were relatively good, 98% - 99.9% considering common and unusual symptoms. Our model outperformed the other experiments [17] that use the same dataset, with a 3-5% increase. We also highlight that unusual symptoms increase the accuracy of the disease prediction task relatively.

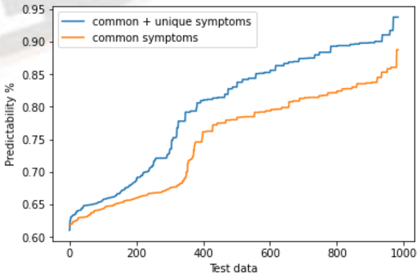

Figure 7. Predictability rate of SVM

Such a result was also validated by estimating the prediction probabilities of the SVM model on the test data, as shown in Figure 7. However, the F1-score of the following diseases is relatively lower, as Acne disease - 81.5%, Impetigo - 87.5%, Paralysis (brain hemorrhage) - 88.8%. The predictability percentage increases by the prediction SVM model considering both common and unusual types of symptoms. However, unusual symptoms were important to diagnose a particular disease, but the F1-score prediction for each disease was relatively high, around 99.9%. Those diseases having a lower F1-score were not correlated to the uniqueness of symptoms and the number of symptoms. The reason could be related to the insufficient dataset. However, understanding common and unusual symptoms are essential in disease prediction tasks; we try to reduce the number of symptoms by employing Principal Component Analysis (PCA). Figure 8 shows the reduced number of symptoms compared to the accuracy of the SVM model. At least four symptoms, regardless of common or unusual characteristics, should be defined for each disease to have a predictability rate of more than 91% by k=5 fold cross-validation.

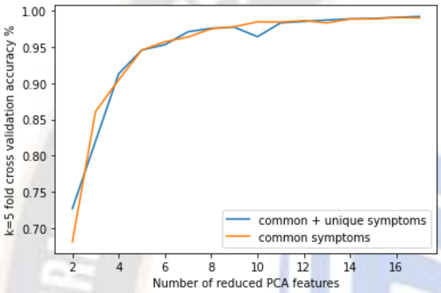

Figure 8. PCA reduced features for SVM prediction

However, significant differences between the predictability considering common symptoms and all symptoms were not observed.

V. CONCLUSION

We conducted a CNN model performance on the extensive amount of MeSH diseases and MeSH symptoms that achieves an accuracy of 97.4% considering the severity of symptoms. In the disease-symptom network, we observed 4,540 nodes and 147,977 edges between diseases and symptoms. We employ linear autoencoders to reduce and extract important features from the MeSH dataset to estimate the efficiency of the model. In this network, 142 symptoms at the maximum are linked to a single disease. Considering at least 84 reduced features out of those maximum 142 features of symptoms linked to a single disease presents more than 90.18% of accuracy on the CNN model which reduces the number of features in the symptoms and also the computation cost of the deep models. Our best achieved performance serves as a baseline model in biomedical analytics. Consequently, we applied the same methodology to a small dataset of clinical notes and estimated the accuracy of 99.9% incorporating both uncommon and common symptoms. Therefore, important features could be reduced to four significant symptoms out of 18 maximum symptoms linked to a single disease with more than 91% of accuracy in the clinical notes.

ACKNOWLEDGMENT

This research was funded by Mongolian Foundation for Science and Technology (MFST) and MES, project No. SHUT-2021/355 "An interdisciplinary network platform for modeling socio-economic and environmental processes in the transboundary territories of the Russian Federation and Mongolia with limited transport accessibility". Also partially supported by RFBR, MECSS, Irkutsk National Research Technical University.

REFERENCES

[1] Zhou, X., Menche, J., Barabási, AL. et al. Human symptoms–disease network. Nat Commun 5, 4212 (2014). https://doi.org/10.1038/ncomms5212

[2] Ting, D.S.W., et al. (2017). "Artificial Intelligence and Deep Learning in Ophthalmology." British Journal of Ophthalmology, 103(2), 167-175.

[3] Luo, G., et al. (2018). "Deep Learning-Based Disease Prediction using Electronic Health Records." IEEE Access, 6, 65333-65341.

[4] Rajkomar, A., et al. (2018). "Scalable and Accurate Deep Learning with Electronic Health Records." npj Digital Medicine, 1, 18.

[5] Choi, E., et al. (2016). "Doctor AI: Predicting Clinical Events via Recurrent Neural Networks." Journal of Biomedical Informatics, 63, 322-327.

[6] Choi, E., et al. (2016). "Retain: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism." In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017-2026.

[7] Lasko, T.A., et al. (2013). "Computational Phenotype Discovery Using Unsupervised Feature Learning over Noisy, Sparse, and Irregular Clinical Data." PLoS ONE, 8(6), e66341.

[8] Kavakiotis, I., et al. (2017). "Machine Learning and Data Mining Methods in Diabetes Research." Computational and Structural Biotechnology Journal, 15, 104-116.

[9] Miotto, R., et al. (2017). "Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records." Scientific Reports, 6, 26094.

[10] Singh, P., et al. (2019). "Disease Prediction Using Data Mining Techniques: A Comprehensive Review." International Journal of Intelligent Systems and Applications, 11(8), 57-68.

[11] Farid, D.M., et al. (2017). "Disease Prediction Using Machine Learning Techniques: A Comparative Study." Journal of Big Data, 4(1), 1-17.

[12] Gholami, M., et al. (2019). "A Comparative Study of Machine Learning Techniques for Disease Prediction." In Proceedings of the International Conference on Artificial Intelligence in Data Science, 97-111.

[13] Naik, G.R., et al. (2020). "A Comparative Study on Machine Learning Techniques for Disease Prediction." Journal of Ambient Intelligence and Humanized Computing, 11(11), 5205-5221.

[14] Mamun, M.A., et al. (2019). "A Review of Machine Learning Techniques for Disease Prediction." In Proceedings of the International Conference on Computing and Communication Systems, 32-36.

[15] Islam, M.M., et al. (2021). "Machine Learning Techniques for Disease Prediction: A Review." International Journal of Advanced Computer Science and Applications, 12(5), 356-364.

[16] Dash, M., et al. (2020). "A Comprehensive Review on Machine Learning Techniques for Disease Prediction." Journal of Ambient Intelligence and Humanized Computing, 11(7), 2943-2966.

[17] S. Grampurohit and C. Sagarnal, "Disease Prediction using Machine Learning Algorithms," 2020 International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-7, doi: 10.1109/INCET49848.2020.9154130.

[18] Dashdorj, Z., Song, M. An application of convolutional neural networks with salient features for relation classification. BMC Bioinformatics 20 (Suppl 10), 244 (2019). https://doi.org/10.1186/s12859-019-2808-3