Medical Text Classification Using Graph Convolutional Network

Abstract—**In this modern era of technology, every day a huge volume of data is being created. Researchers are using cutting-edge technology to understand the data for extracting useful insights. In the field of data science and artificial intelligence, these data are playing a significant role to contribute to the advancement of technology. Using these data algorithms, like deep learning, machine learning, and natural language processing, are used to solve advanced problems, like object recognition, business prediction, and analysis. Among these graph-related problems like traffic forecasting, image classification and social media graph analysis, are growing every day. We can find graphs all over the world. The internet itself is a graph network of devices connected to each other, along with social media, molecules etc. In our research, we are proposing a way, which takes texts of medical symptoms and classifies them to the kind of medical branches, like dermatologist or dentists, the text fits best, using Graph Convolutional Network (GCN). In this proposal we have aimed to find the connection between the nodes (patients) which have been classified to the same type of medical branch from the texts of symptoms or past history of their medical records. After applying a GCN approach we have come to a result which can classify 90.2% medical text correctly to their needed specific medical branches.**

Index Terms—**Graph Convolutional Network, Natural Language Processing, Deep Learning, Text classification**

I. Introduction

Text classification is a field of immense possibilities and applications in natural language processing (NLP). Classifying customers' reviews, medical text, social media comments, and posts are some of the applications of text classification. In general, text classification means extracting information from the given text and classifying it into its labeled category. Most of the past approaches to counting NLP applications generally depended on TF-IDF scoring [2,3]. Later, machine learning algorithms like SVM, ANN, etc started performing well in classifying the sentiment of texts [4]. After that, researchers used deep learning techniques to achieve better accuracy. For text classification, mostly deep learning techniques are used. However, recently Graph Neural Networks are outperforming traditional deep learning techniques. In this modern era of technology, a huge volume of electronic medical records and literature is being generated every day. A good amount of medical data is available online, and this data provides useful information about the disease, symptoms, treatment, patient history, medication, and so on. To imbibe the most useful information, they need to be classified into their respective classes [1]. The major contribution of our paper is that, we are using GCN which uses neural network inside a graph architecture, which better at classifying text, than any multi layer perceptron (MLP) or support vector machine (SVM), for texts related to medical symptoms.

II. Related Work

Masoud Malekzadeh et al. [5] said that natural language processing (NLP) has many applications in a number of sectors, including social networking sites, healthcare, and biochemistry. The bag-of-words model is one of the simplest methods for representing real language. A more sophisticated model takes into account the relationships between the words in a text in addition to their frequencies. The model considers the token order within a sentence and within a text.According to experimental findings, ST-GCN outperformed SWEM, the second-best baseline model, in accuracy for the Amazon Internal dataset by a margin of 5.86%. A more sophisticated model takes into account the relationships between the words in a text in addition to their frequencies.The model considers the token order within a sentence and within a text. They have outperformed the traditional and deep learning based equivalent on the reported data sets, according to their evaluation results. Yutai Duan et al. [6] proposed Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are two examples of deep learning neural networks that have recently found success. These networks are mostly used in pattern identification and data mining to identify hidden patterns in Euclidean data. They Classified Convolutional GNN frameworks come in two varieties: node classification and graph classification. In order to achieve this, they created a GRLB to learn the entire node representation and incorporate new levels of information. Next, we created a 1-hop SAPooling layer to create a representational subgraph structure of roots. Finally,They performed RCGNN ablation experiments, which demonstrated the efficiency and understandability of our modules. Yaohui Hu* et al. [7] described about the Urban Road Traffic System in his paper. Urban traffic congestion is growing increasingly crowded, and transport effectiveness is poor, which significantly limits the city's ability to develop healthily. Traffic data is gathered by placing sensors along the roadside, which is necessary for deep learning approaches to anticipate traffic flow. This data format may be processed and the features from the data can be extracted effectively using convolutional neural networks. First they Apply a graph convolutional network to capture the traffic flow, treat the sensor data simultaneously as a graph, connect the nodes and neighbor nodes to reflect the spatial correlation of the traffic flow. Then they consider the data for various times and analyze the consideration factors of the traffic flow. Thus by creating a graph, this study successfully predicts urban traffic flow. Zhao-Min Chen et al. [8] proposed a multi-labeled image recognition to predict a set of object labels that present in the image. Then they proposed a multi-labeled classification model based on Graph Convolutional Network (GCN). The process builds a directed graph over the object labels, where each node is represented by a word embedding of the label and GCN learns this model graph into a set of inter-dependent object classifiers.Yongjian Ren et al. [9] predict a model that medical treatment migration based on medical insurance data is introduced in their paper. Medical treatment graph is constructed based on medical insurance data. This medical treatment graph is a heterogeneous graph which contains diseases, hospitals, medicines, hospitalisation events and the relation between these entities. Mark Cheung et al. [10] proposed that recent research has found that using graph neural network (GNN) models has given enough data which can perform better than using human-engineered fingerprints or descriptors in predicting molecular properties of potential antibiotics. Youngchul Kwak et al. [11] proposed a system where the brain-computer interface (BCI) system provides information exchanges between neural signals containing the user's intention and device control signals. They present a graph neural network (GNN) with a multilevel feature fusion structure for high performance BCI systems. Ahmad Al-Doulat et al. [12] proposed using deep learning techniques and linguistic analysis (Semantic and Statistical) techniques for unstructured medical text classification at the document level which handles multiclasses medical articles classification. To execute this structure they used two types of features: (i) content-based features (stylistic and complexity), (ii) health domain-specific features and also applied NLTK in preprocessing steps. Their preprocessing procedure contains a number of processes, including text sanitization, stop words/punctuation removal, sentence splitting, POS tagging, word tokenization using, word lemmatization and Named Entity Recognition (NER). Their resulting accuracy is 82%. They also utilized the-state-of-the-art TF/IDF for this purpose and acquired 62% accuracy result. Thomas N. Kipf et al. [13] present a scalable method for semi-supervised learning on graph-structured data which is based on an convolutional neural network variant that operates directly on graphs, scales linearly in the number of graph edges, and learns hidden layer representations that encode both local graph structure and node features. In this structure they used TensorFlow for an efficient GPU-based implementation and considered three citation network datasets: Citeseer, Cora and Pubmed which contain a list of citation linkages between papers as well as sparse bag-of-words feature vectors for each document. CHENSHENG LI et al. [14] proposed a scalable graph convolutional neural network with fast localized convolution operators derived from directed graph Laplacian, known as the fast directed graph convolutional network (FDGCN) and it can directly work on directed graphs and scale to large graphs as the convolution operation is linear with the number of edges.

III. Methodology

A. Worflow Diagram

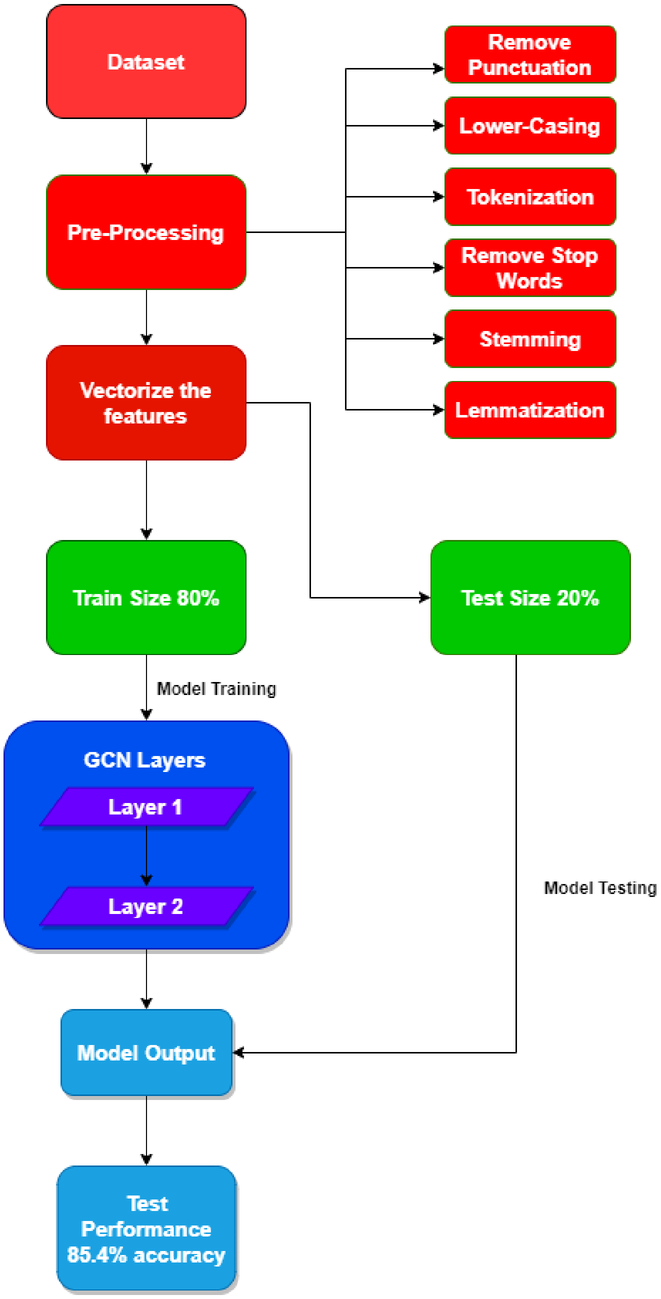


Fig. 1. Workflow Diagram

B. Dataset

Dataset was collected from the Kaggle website [16]. The dataset has 4999 data. In the dataset "description" columns were the supposed training data which have text that are to be classified. And the column "medical_specialty" has a value of 40 unique medical branches which are to be classified as labels from the text column data. The dataset was divided into a training dataset of 3999 rows (80%) and a test dataset of 1000 rows (20%) as shown in fourth step Figure 1, after preprocessing. Figure 2 shows the bar plot summary of the dataset rows according to 40 labels.

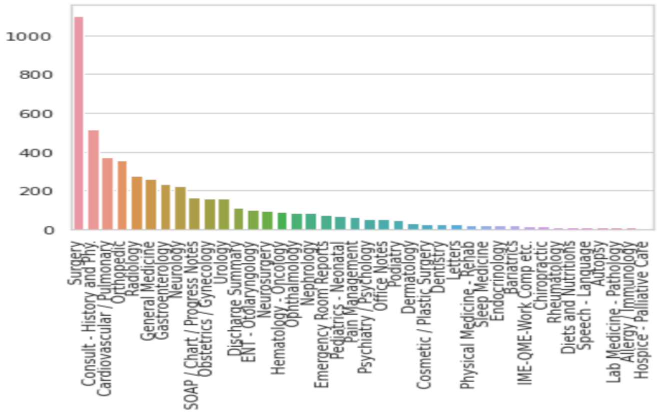


Fig. 2. Dataset Summary

C. Data pre-processing

The texts are first Tokenized to form the corpus of important words and then lemmatization, stemming processes were applied to the corpus for better and more effective corpus in the first step of preprocessing. These processes make the corpus unambiguous and more effective in processing for training or testing, by converting similar words with the same meaning to only one word and also removing stopwords from the corpus. The Python tool of Spacy was used for pre-processing. Thus these corpus has less encoded messages which are passed along for the next step of feature extraction. In the feature extraction step, the text messages and the corpus was used to vectorize the tokens in the text messages, to form the training and testing data and passed to the GCN layers.

D. Graph Convolutional Network

$$\mathbf{X}^{(\ell+1)} = \sigma\left(\tilde{\mathbf{A}}\mathbf{X}^{(\ell)}\mathbf{W}^{(\ell)}\right)$$

Fig. 3. GCN Equation

According to [10] Graph Convolutional Network is a Semi-Supervised Classification technique which uses the concept of graphs of vertices and edges to classify the nodes with similar traits to be connected more closely to classify. GCN uses graph of G = (V, E) where V is the number of vertices and E is the number of edges in the adjacency matrix A ϵ R^{N*N}. So, using each person as a vertex and the connection of the vertices and the medical branches.

Figure 3 is the general formula of a GCN [15]. Ā is the adjacency matrix Ā = A + IN and W is the weight matrix of the along with a sigmoid function for the hidden layer propagation in the nodes. X is the node feature.

In our research we have used a two-layered GCN shown in Figure 4 as an example. In the first layer, each node has the given features vectors from the input data which are vectorized in the pre-processing step. In the first layer each text of the dataset is a node of the graph where the node features are the vectorized data done in the previous step. The input node representations are processed using a FFN to produce a message. Apply preprocessing using feed forward network (FFN), which has 2 deep hidden layers like given in Figure 5 in each node, to the node features (messages) to generate initial node representations. Then messages of the neighbors of each node are aggregated with respect to the edge weights given in the training step, and then using a permutation invariant of sum pooling operation to the aggregated message. The node representations and aggregated messages are combined and processed to produce the new state of the node representations (node embeddings). Then the node embeddings are then sent to the next convolution layer with skip connections, which repeat the same process to the node representation to produce node embeddings. After that applying a post-processing using FFN to the node embeddings to generate the final node embeddings. Finally, we feed the node embeddings in a Softmax layer to predict the node class

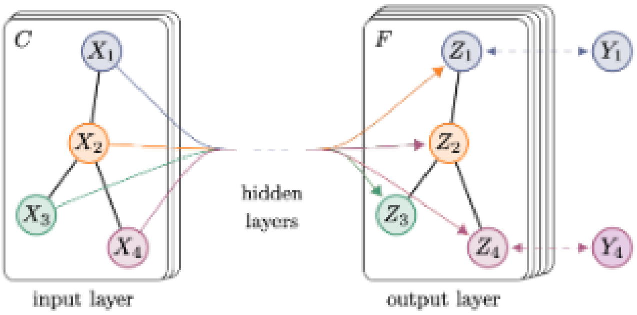


Fig. 4. Convolution Layers

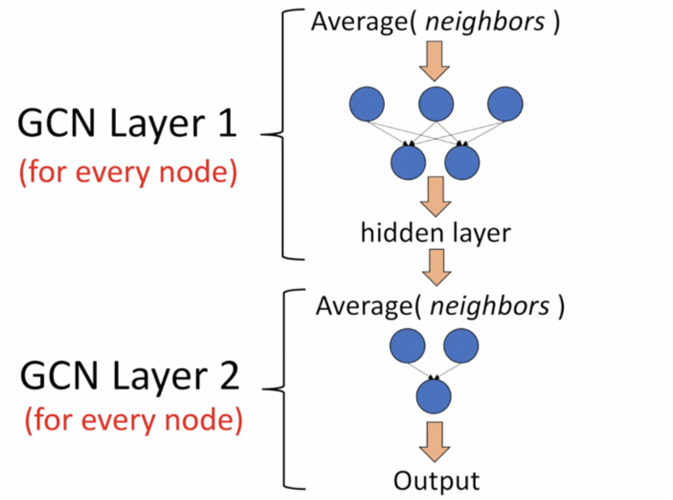


Fig. 5. Hidden Layer in the nodes

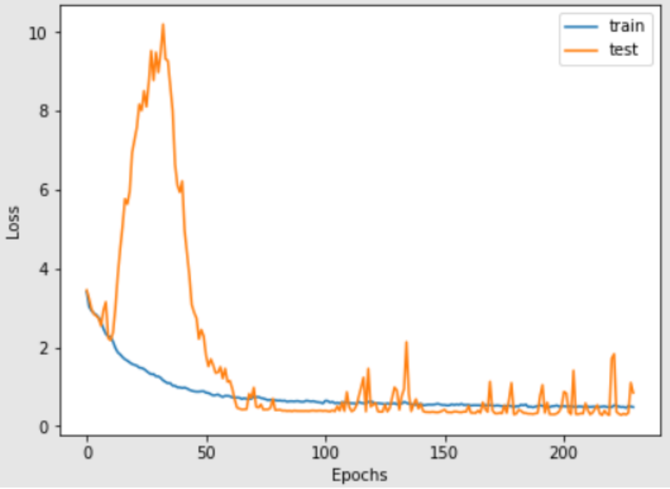


Fig. 6. Training Loss Graph

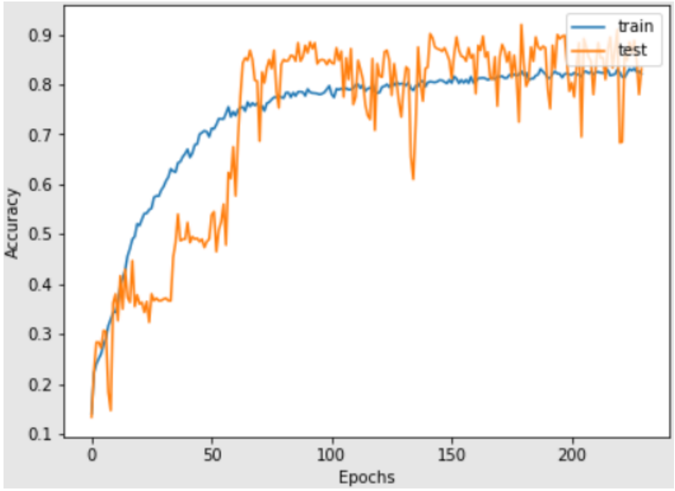


Fig. 7. Training Accuracy Graph

IV. Result and Analysis

A. Result

Figure 7 represents the training and validation accuracy of 300 training epochs, and maximum accuracy model was collected to be trained on untrained test dataset, and average accuracy of 90.2% was obtained in the GCN model which is better and more precise performing than any other model.

B. Analysis

As shown in the Figure 9 we have tried various models like support vector machine (SVM), multi layer perceptron (MLP) and graph convolutional network (GCN) and as it can be seen that GCN is performing the best with highest precision(88%), recall(90%) and f1-score(88%) as well. Figure 8 shows the heat map matrix of predicted values of GCN model and it clearly shows the diagonal have high similarity between predicted and original data on our proposed model. This shows that GCN can classify by nodes more accurately than other linear or non-linear models and perform more efficiently. Many paper shown accuracy of 82% using models like artificial neural network, our model outperform those [12]. Although there are papers which have achieved higher accuracy, our model was also close and with further fine tuning and hybrid methods it can perform better.

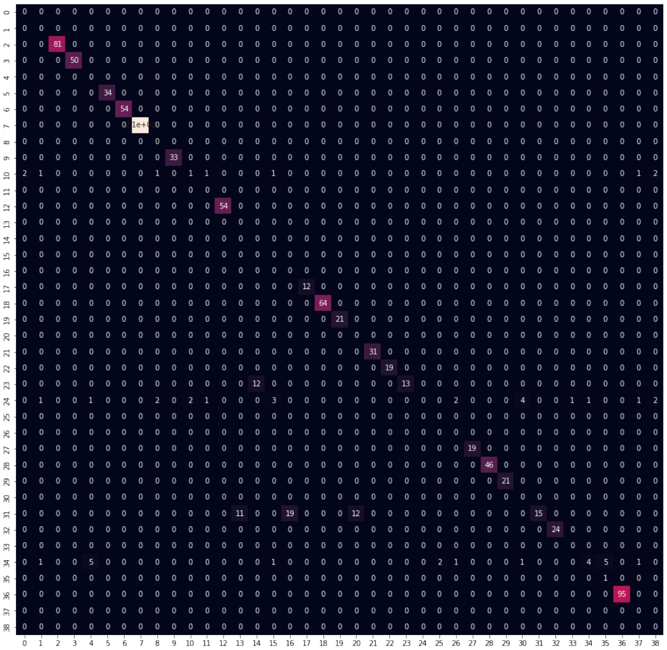
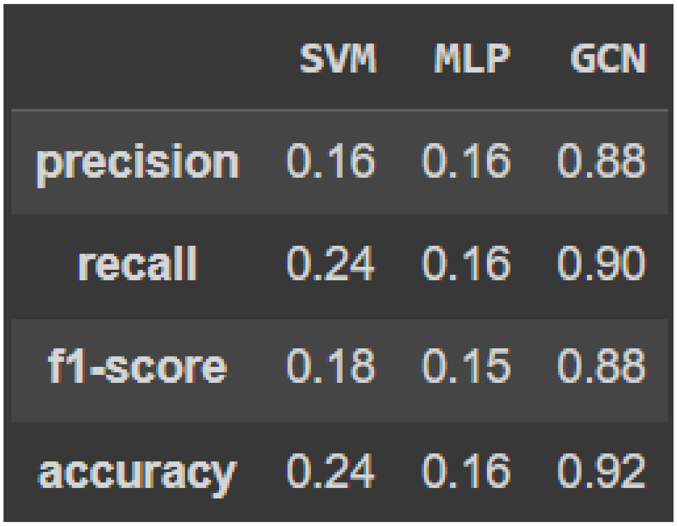


Fig. 8. GCN confusion matrix



	SVM	MLP	GCN
precision	0.16	0.16	0.88
recall	0.24	0.16	0.90
f1-score	0.18	0.15	0.88
accuracy	0.24	0.16	0.92

Fig. 9. Model Comparison

V. Conclusion

In this paper, we have put an effort into classifying the medical transcription data. The used Graph Convolution Network works very well in this regard and does much better work in text classification than many other researches using only two layers. Transformer models might also work very well in these kinds. Finally, it paves a new way for future research in medical text classification.

References

[1] S. K. Prabhakar and D.-O. Won, "Medical Text Classification Using Hybrid Deep Learning Models with Multihead Attention," Comput. Intell. Neurosci., vol. 2021, pp. 1–16, Sep. 2021, doi: 10.1155/2021/9425655.

[2] G. Groh and J. Hauffa, "Characterizing Social Relations Via NLPBased Sentiment Analysis," Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, pp. 502-505, Barcelona, Catalonia, Spain, 2011.

[3] A Balkiwal, P. Arora, A. Patil, and V. Varma, "Towards Enhanced Opinion Classification using NLP Techniques," Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP), pp. 101-107, 2011.

[4] R. Moraes, J. F. Valiati, and W. P. G. Neto, "Document-level Sentiment Classification: An Empirical Comparison between SVM and ANN," Expert Systems with Applications, vol. 40, no. 2, pp. 621-633, 2013.

[5] M. Malekzadeh, P. Hajibabaee, M. Heidari, S. Zad, O. Uzuner, and J. H. Jones, "Review of Graph Neural Network in Text Classification," in 2021 IEEE 12th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), New York, NY, USA, Dec. 2021, pp. 0084–0091. doi: 10.1109/UEMCON53757.2021.9666633.

[6] Y. Duan, J. Wang, H. Ma, and Y. Sun, "Residual convolutional graph neural network with subgraph attention pooling," Tsinghua Sci. Technol., vol. 27, no. 4, pp. 653–663, Aug. 2022, doi: 10.26599/TST.2021.9010058.

[7] Y. Hu, "Research on City Traffic Flow Forecast Based on Graph Convolutional Neural Network," in 2021 IEEE 2nd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE), Nanchang, China, Mar. 2021, pp. 269–273. doi: 10.1109/ICBAIE52039.2021.9389951.

[8]]Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-Label Image Recognition with Graph Convolutional Networks." arXiv, Apr. 07, 2019. Accessed: Aug. 30, 2022. [Online]. Available: http://arxiv.org/abs/1904.03582

[9] Y. Ren, Y. Shi, K. Zhang, Z. Chen, and Z. Yan, "Medical Treatment Migration Prediction Based on GCN via Medical Insurance Data," IEEE J. Biomed. Health Inform., vol. 24, no. 9, pp. 2516–2522, Sep. 2020, doi: 10.1109/JBHI.2020.3008493.

[10] L. Yao, C. Mao, and Y. Luo, "Graph Convolutional Networks for Text Classification." arXiv, Nov. 13, 2018. Accessed: Aug. 30, 2022. [Online]. Available: http://arxiv.org/abs/1809.05679

[11] Y. Kwak, W.-J. Song, and S.-E. Kim, "Graph Neural Network with Multilevel Feature Fusion for EEG based Brain-Computer Interface," in 2020 IEEE International Conference on Consumer Electronics - Asia (ICCE-Asia), Seoul, Korea (South), Nov. 2020, pp. 1–3. doi: 10.1109/ICCE-Asia49877.2020.9276983.

[12] A. Al-Doulat, I. Obaidat, and M. Lee, "Unstructured Medical Text Classification using Linguistic Analysis: A Supervised Deep Learning Approach," in 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA), Abu Dhabi, United Arab Emirates, Nov. 2019, pp. 1–7. doi: 10.1109/AICCSA47632.2019.9035282.

[13] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks." arXiv, Feb. 22, 2017. Accessed: Aug. 30, 2022. [Online]. Available: http://arxiv.org/abs/1609.02907

[14] C. Li, X. Qin, X. Xu, D. Yang, and G. Wei, "Scalable Graph Convolutional Networks With Fast Localized Spectral Filter for Directed Graphs," IEEE Access, vol. 8, pp. 105634–105644, 2020, doi: 10.1109/ACCESS.2020.2999520.

[15] . M. Cheung and J. M. F. Moura, "Graph Neural Networks for COVID-19 Drug Discovery," in 2020 IEEE International Conference on Big Data (Big Data), Atlanta, GA, USA, Dec. 2020, pp. 5646–5648. doi: 10.1109/BigData50022.2020.9378164.

[16] Boyle, T. (2018, September). Medical Transcription, Version 1. Retrieved 20 July, 2022 from https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions.