EVALUATION OF MULTIVARIATE DATA ACQUISITION OF NETWORK EMBEDDING SCHEME FOR HEALTHCARE APPLICATIONS

PAVAN MOHAN NEELAMRAJU[1], MOHAMED EL-DOSUKY[2,3], SHERIF KAMEL[2,4]

[1]Next Tech Lab, SRM University - AP, Andhra Pradesh, India.

[2]Computer Science Department, Arab East Colleges, Saudi Arabia

[3]Computer Science Department, Faculty of Computers and Information, Mansoura University, Egypt

[4]Department of Communications and Computer Engineering, October University for Modern Sciences and Arts, Egypt

E-mail: npavanmohan3@gmail.com, maldosuky@arabeast.edu.sa, skhussein@arabeast.edu.sa

ABSTRACT

In recent years, recommendation systems have evolved to provide valuable information with respect to a multitude of domains. In the ongoing medical revolution, they have been widely utilized for identifying trends from electronic record data. Techniques were developed to compute the correlation between patients suffering from diseases with similar symptoms, identification of treatment procedures and drug identification. However, the feasibility of the heterogeneous network embedding schemes needs to be studied in the dimension of varying input data formats and the corresponding performance. In the following work, emphasis is placed on understanding the impact of a variety of data, volume and the number of recommendations produced by the system. Further, the metrics such as precision and recall were utilised to evaluate the overall performance of the recommendation system, which is built upon Metapath2Vec. The current study extrapolates the effectiveness of the recommendation system using data gathered from 1500 influenza patients, further elucidating the ability of recommendation systems to identify distinct trends from the disease's symptoms that are perceptible to people. In addition to the implementation of Metapath2Vec and corresponding analysis, a detailed note is provided on elucidating the future of network embedding schemes and recommendation systems.

Keywords: *Network Embedding, Recommendation System, Metapath2Vec*

1. INTRODUCTION

Due to the increase in population, the need for technological innovation in the healthcare sector has been more important than ever. The current technological innovations led to a drastic impact on the frontiers of providing accessible monitoring systems, efficient disease tagging, imaging methods and cost-effective diagnostic procedures. However, the need for a human supervisor to provide more personalized and case-by-case diagnoses has been a standalone procedure. The manual diagnostic process is a culmination of various steps such as patient identification, description of symptoms, comparison with reference information and cause-based treatment.

In the following flow, the physician examines a multitude of attributes concerning the present state of the patient, the nature of the health condition, causation factors and possible therapeutic strategies. The data used during the aforementioned steps are of different types and formats. This leads to reduced information gain and the possibility of data being ambiguous. Furthermore, it becomes difficult for a physician to integrate the heterogeneous data collected to meet the requirements. In addition to the same, different people perceive the same symptom with varied intensities, making the identification of particular diseases difficult.

The problem of data integration and decision-making can be ameliorated by mapping the relationship between various attributes and the outcome. Converting the following problem into a computer paradigm requires a data structure that can describe the relations and interactions in the pool of features. On similar lines, by considering various

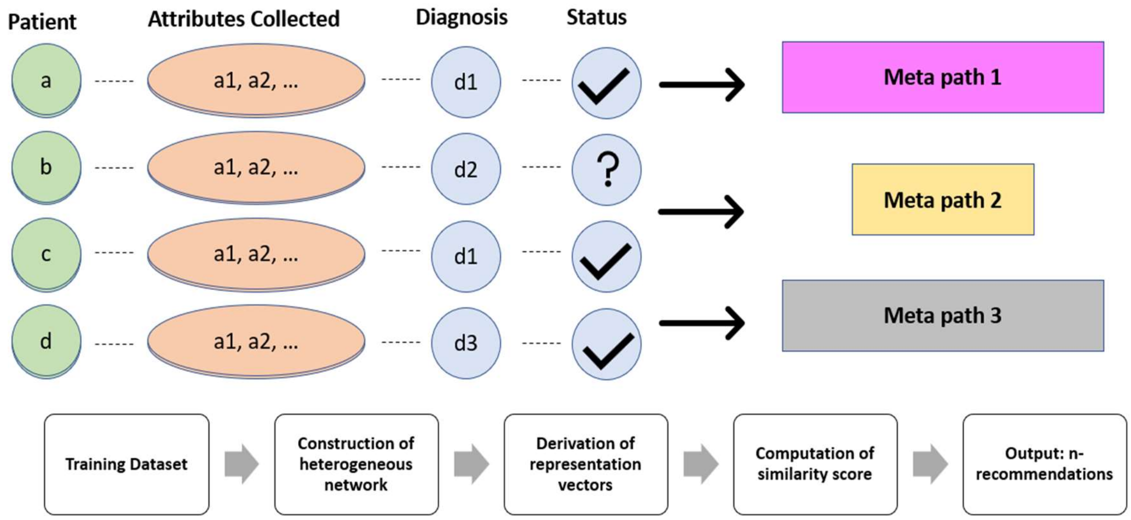
Patient
Attributes Collected
Diagnosis
Status
a
b
c
d
a1, a2, ...
d1
d2
d1
d3
✓
?
Meta path 1
Meta path 2
Meta path 3
Training Dataset
Construction of heterogeneous network
Derivation of representation vectors
Computation of similarity score
Output: n-recommendations

Figure 1: Illustration of Metapath2Vec process flow

attributes as nodes and the relationship between them as the edges, it is possible to construct a heterogeneous graph that contains different node types. Heterogeneous graphs enable more possibilities for accessing the multidimensional interaction between its nodes. Albeit the multidimensional interactions of the graph are often constrained by the diversity of nodes that exist within the network. For an instance, few relations may only occur between two specific node types [1].

The utilization of heterogeneous networks and their amalgamation with Machine Learning techniques can drastically improve diagnostic procedures [2]. Further, this enables smarter models that can work on historical data to examine the underlying patterns. Along with the same, in comparison to conventional ML algorithms, network-based approaches provide valuable insights regarding connections between individual data points and their properties.

The current work encompasses the procedure of node embedding to aggregate the position of an individual node, its position, and its local neighbours. The following process entails the generation of an encoded feature vector named, node embedding.

Node embedding helps in scrutinizing the node properties and the relations that exist between a node and its local neighbours. The attained feature vector is further used as an input to a recommendation system as an additional feature. Therefore, the node embedding procedure enhances the efficacy of a learning system by extrapolating the hidden structural information and trends about the data points, that was unless not observed with a simple input vector [3].

The current paper provides a concise summary of the previous literature and thereby, provides an emphasis on the merits of the proposed system and its implementation in the domain of healthcare in Section 3. Furthermore, the results are presented in Section 4 after a full discussion of the modality's data collection, implementation, and development.

2. BACKGROUND

The viability of using network embedding techniques in the field of medical diagnostics has been the subject of several research studies and analyses in the past. In particular, the literature review can be viewed as a concatenation of three preliminary research dimensions. Namely, research was conducted in the facets of drug prediction, development of novel algorithms and healthcare monitoring frameworks.

A novel heterogeneous network embedding technique is proposed to learn the drug representations and to integrate the protein-to-protein interactions to predict adverse drug reactions. Three different network embedding models were assessed against their efficacy in predicting adverse drug reactions and it is reported that Signed Heterogeneous Information Network Embedding (SHINE) is more effective [4]. On the other hand, heterogeneous weighted networks also found their use case in the identification of effective herbs in

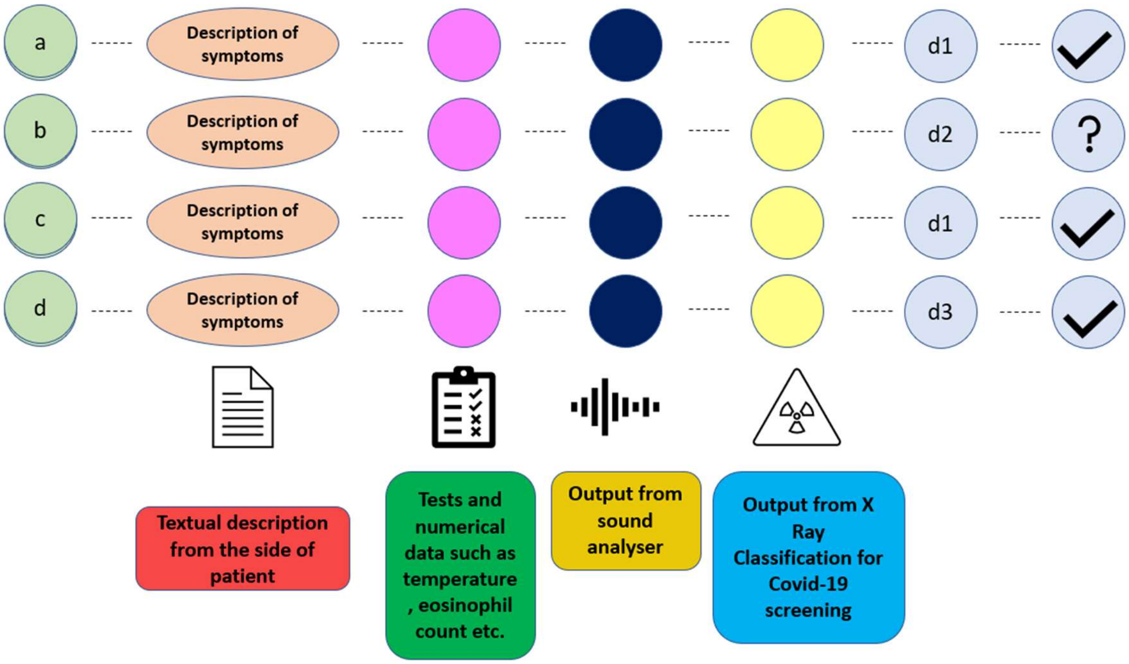
a
b
c
d
Description of symptoms
Description of symptoms
Description of symptoms
Description of symptoms
d1
d2
d1
d3
✓
?
✓
✓
Textual description from the side of patient
Tests and numerical data such as temperature , eosinophil count etc.
Output from sound analyser
Output from X Ray Classification for Covid-19 screening

Figure 2: Aggregation of multiple data sources

traditional Chinese medicine. An unsupervised analysis model is proposed to realize the same [5].

In addition to the aforementioned research works, Graph Neural Networks and Bidirectional Encoder Representation from Transformers (BERT) were utilized for medication recommendation and medical code representation. A new model of Graph-BERT (G-BERT) is proposed to address the selection bias [6]. Consecutively, these representation-based learning algorithms found their implementations in biomedicine and healthcare [7]. Trends, techniques, and applications of graph representation and learning schemes are elaborated [8].

Along similar lines, multiple research works have been done in the dimension of healthcare management. A multitask recommendation system is developed depending on deep neural networks and 5G networks. The developed modality was validated in terms of treatment recommendation and disease prediction [9].

Likewise, social networking platforms and wearable sensors were used to collect patient data for monitoring healthcare. To overcome the challenges in handling big data, a novel monitoring framework is developed depending on the big data analytics engine and cloud environment. The system was used for the classification of blood pressure, mental health, diabetes status and drug reviews [10]. The similarity between the users in heterogeneous networks is utilized to construct a user recommendation system. Global and local structure methods were presented to compute the similarity [11].

In the paradigm of healthcare monitoring frameworks, several other studies extrapolated important insights such as electronic health record mining [12], predicting effective diagnostic procedures [13] and medication recommendations [14]. Contrary to the symptom focussed recommendation, studies were done to identify the right physician depending on disease-specific expertise. In the prior study, two different embeddings are fused to connect a doctor and patient pair, which are mapped to one another. The created model is used to address the ill effects of group heterogeneity and data imbalance, by simultaneously generating expertise scores for doctors [15].

Apart from multiple features and relations as discussed above, various methods were put forward to extrapolate the trends within the patients experiencing similar health conditions. Dynamic Bayesian Embedded Recurrent Neural Networks are used to learn fine-grained patient similarity. The model is used to determine the distance among patients and causal correlations between various medical events [16]. Similarly, progress was made in the direction of algorithmic development [17-19].

From the above-discussed literature, it can be concluded that though there has been significant progress in the direction of drug prediction, healthcare management and the development of new algorithms, there is enough scope for the development of new techniques that can extrapolate meaningful information from a multitude of data formats and their corresponding evaluation. The current study focussed on evaluating one such technique of Metapath2Vec concerning a variety of data, volume and the count of recommendations generated, thus, paving way for an all-around patient diagnosis and confident characterization of the medical conditions.

3. METHODOLOGY

3.1 Preliminary Analysis

In the following paper, we examine the ability of heterogeneous networks to articulate the data collected from various sources (data variety) and recommend the best possible diagnosis depending on the electronic record data. The current study further elucidates the capability of recommendation systems to recognise different patterns from the disease's symptoms that are perceptible to individuals by extrapolating the efficiency of the recommendation system using data collected from 1500 influenza patients. The dataset contains a list of symptoms experienced by individual patients and the corresponding test results that were conducted on them. The details of the patients were such as name and hospital enrolment ID were removed. Instead, patient ID was generated corresponding to the patient enrolled to maintain the anonymity, which is further used as a primary key for a particular sample.

By considering the symptoms of the patients as attributes, we provide similar patient case studies to make the diagnosis easier. There are five main stages in the creation of the following model. The initial step involves data preparation and cleaning. The second step involves the creation of a heterogeneous information network from the data that was utilized, wherein, the network can be considered a set of nodes, edges, and relations. Further, the third step involves the creation of multi-relation paths. Consequently, meta-path-based random walkers were utilized to generate the paths for learning, followed by the implementation of Metapath2Vec. After the model-building steps mentioned above, a similar patient case study recommendation is done, where input data is collected from various decision-making processes utilizing data formats such as images, numerical test records and text. The entire process is summarized in Fig. 1.

Experiences collected from the side of the patient i.e., their symptoms come under the textual information. Whereas their lung status from X-ray scans and sound analysers constitutes the information collected from the formats of images and signals respectively.

The training data set could be considered a discrete set of various parameters, explaining the status of the patient (P) and the corresponding attribute information such as keywords from the textual description $\{t_1, t_2 .. t_n\}$, results from tests $\{r_1, r_2, r_3\}$ (body temperature, eosinophil count and CBNAAT result), metrics of sound analyser frequency $\{f\}$ and the X-Ray based screening for Covid-19 screening $\{x\}$.

Therefore, the training set can be observed to be a concatenation of various attributes as a vector. The representation of the same is observed to be as follows,

$$P = \{t_1, t_2, \dots, t_n, r_1, r_2, r_3, f, x\} \qquad (1)$$

Simply put, the heterogeneous network (G) is simply a mathematical formulation representing the relation between the nodes or vertices (V), edges (E) and the relationship (P), represented as

$$G = \{V, E, P\} \qquad (2)$$

Further, with the object type mapping and relation type mapping functions in eq (3) and eq (4) respectively, a heterogeneous information network is associated. The set (A_V, A_E) denotes the network meta, which acts as a defining entity for the heterogeneous network.

$$f(v): V \to A_V \qquad (3)$$

$$\Psi(e): E \to A_E \qquad (4)$$

3.2 Model Construction

The attributes as stated above, include a unique patient identification number, their corresponding symptoms such as body temperature, cough, sore throat, runny nose, muscle ache, headache, fatigue, diarrhoea, tests done, the mode of treatment given and the outcome of the diagnosis etc. Along with the same, the results of various tests such as acoustic pulmonary test, and blood test is also present, illustrating the lung condition and eosinophil count.

As discussed above the current study uses a culmination of various decision-making paradigms providing various information, this includes a deep learning-based respiratory sound analyser [20] and X-ray classification for Covid-19 screening [21], which works upon acoustic signals and X-ray images respectively. The aggregation of various data sources is illustrated in Fig. 2.

Additionally, the suggested CN_HER technique [22] is used in the subsequent study to gather and project the essential data on heterogeneous networks in order to provide comparable patient-case recommendations. Therefore, the problem of creating a recommendation system can be articulated as a transformation task, where heterogeneous information network embeddings are used to adapt to the low-dimensional embeddings.

Further, the adaptation with respect to the embeddings is used to capture the characteristics of the heterogeneous data. Furthermore, the implementation of CN_HER demonstrates its ability to capture the trends from heterogeneous data. Simultaneously this acts as an alternative to the existing homogeneous network-based techniques such as a random walk to initiate node sequences, as in the case of online learning of social representations [23]. Nodes and edges with various object and relation categories cannot be contradistinguished using the aforementioned paradigm. The Heterogeneous Information Network Embeddings must thus be traversed in a more flexible manner in order to provide significant node sequences.

Along with the same, the current study utilizes a majority of data in the text format which contains patients' experiences and descriptions of symptoms covering various keywords such as cough, sore throat, runny nose, muscle ache, headache, fatigue, and diarrhoea. The aforementioned attributes are extracted from the text provided by the patients studied. Therefore, in order to grasp the following attributes as significant symptoms, it is pivotal to generate effective node sequences. To establish the same, meta-paths are employed to understand the semantic trends of heterogeneous network embeddings. On similar lines, a meta-path-based random walk technique is used. In the following method, electronically collected patient data is provided as the input which contains the previously mentioned symptoms. Additionally, the diagnosis provided, and the current status of the patient were also present. Concurrently, the input also contains the meta-path scheme, walk length, vector dimension, size of the neighbourhood and test data for validation of the constructed model (used in the final phase for verification).

The model is expected to construct a heterogeneous information network corresponding to the medical patient record. Various representation vectors were obtained using Metapath2Vec, after which, the training patient cases are ordered according to their respective similarity scores between the input patient case and other patient cases in the dataset.

To conclude the methodology section, the input to the recommendation system is the sequence of words of the training patient cases and testing patient cases with the respective symptoms, diagnosis given and the patient status, along with the adjacency matrix derived based on the network. Thus, various vector representations were used to calculate the similarity scores. Finally, the training patient cases were ordered in accordance with the computed similarity scores and the patient cases with high similarity scores were considered for the final recommendation.

Hence, the implementation of the current recommendation system is a culmination of various procedures that try to extract the similarity between the patient records, which are mathematically modelled. The next section of the paper deals with the experimental validation of the current algorithm in extrapolating meaningful information from the obtained patient record.

4. RESULTS

After understanding the implementation of the following algorithm in articulating the needed similarity within the patient group, it is important to compare it with respect to some other methods that have been widely utilized. In addition to the training data provided, 500 additional samples of data related to patients infected with influenza are utilized for validating the implemented model. The textual description corresponding to each patient case is extracted. However, even the validation dataset is completely processed and involves- extraction of patient symptoms from the respective textual

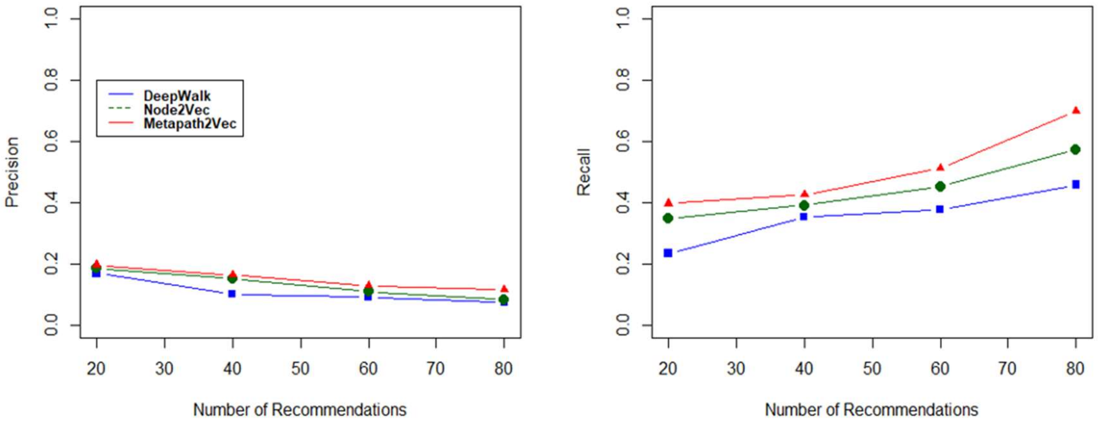

Figure 3: Variation of precision and recall across the number of recommendations

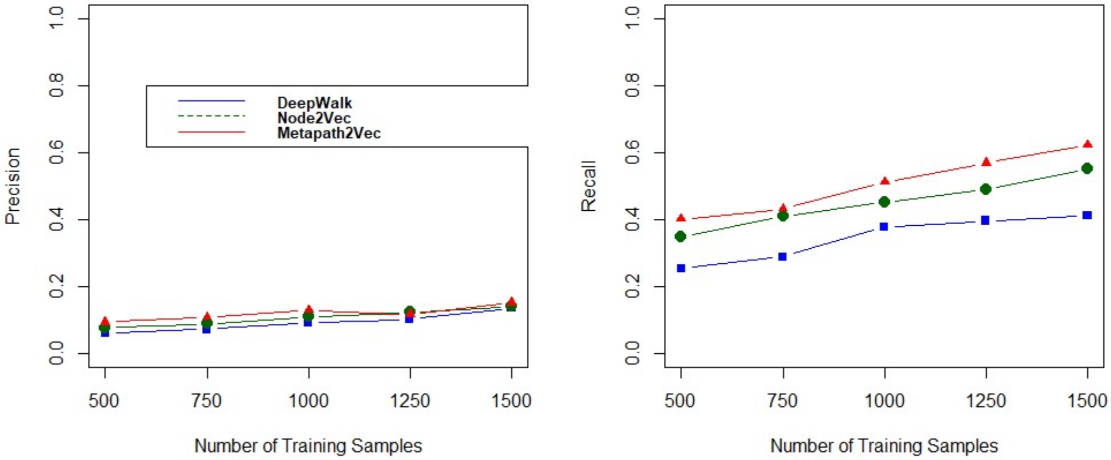

Figure 4: Variation of precision and recall across the number of training samples

description, concentrating on words that are of greater significance and removal of redundant stop words for ensuring efficient symptom capture. After pre-processing the necessary data, the final step involves the recognition of keywords (corresponding to the patient symptom), which further helps in the construction of the patient-symptom network. To realize the same, the term frequency and inverse document frequency (TF-IDF) approach is implemented. A list of 47 words was identified as possible symptoms by selecting 0:03 as a threshold limit.

Finally, for validating the current model two different metrics of precision and recall were employed. The following metrics were utilized to assess the implementation of the model in capturing the trends from the testing data provided. Similar to the training dataset, the patient ID has been the sole primary key in tagging a particular observation. The aforementioned metrics were used to compare the efficacy of Metapath2Vec with two similar graph representation approaches DeepWalk and Node2Vec.

Working upon the recommendations produced by DeepWalk and Node2Vec, the performance is compared with that of Metapath2Vec, by considering the number of required recommendations as 20, 40, 60, and 80, by selecting

Table 1: Recorded Precision and Recall with respect to number of recommendations

Method	Number of Recommendations	Precision	Recall
DeepWalk	20	0.169	0.234
	40	0.101	0.354
	60	0.092	0.378
	80	0.076	0.458
Node2Vec	20	0.186	0.347
	40	0.152	0.391
	60	0.109	0.452
	80	0.083	0.573
Metapath2Vec	20	0.197	0.396
	40	0.163	0.425
	60	0.129	0.512
	80	0.116	0.698

Table 2: Recorded Precision and Recall with respect to the training data size

Method	Number of Recommendations	Precision	Recall
DeepWalk	20	0.169	0.234
	40	0.101	0.354
	60	0.092	0.378
	80	0.076	0.458
Node2Vec	20	0.186	0.347
	40	0.152	0.391
	60	0.109	0.452
	80	0.083	0.573
Metapath2Vec	20	0.197	0.396
	40	0.163	0.425
	60	0.129	0.512
	80	0.116	0.698

1000 training samples. Their respective performances were compared in Fig. 3 depending upon the aforementioned metrics.

Consequently, the final component of the evaluation is done concerning the training data size, wherein the volume of data is varied across 1500, 1250, 1000, 750 and 500 samples. The respective precision and recall were shown in Fig. 4. considering the number of recommendations as 60. For the purposes of better understanding, the information provided in Fig. 3. and Fig. 4. are tabulated in Table – I and Table – II. The entire process cycle of methodology and validation is illustrated in Fig. 5.

5. CONCLUSION

The premise of the paper can be looked at in two different dimensions. Firstly, the following work examines the efficacy of Metapath2Vec, Node2Vec and DeepWalk techniques in articulating the information from the data provided. The primary objective of the following work lies in examining the relevance of recommendation systems in the field of medical diagnosis and their respective abilities to provide the proportion of relevant recommendations, along with the proportion of recommendations that are relevant. However, the novelty of the following work relies upon how effectively the current model understands the information from a patient's point of view, thus illustrating the model's ability to capture the humanistic paradigm of semantic and structural information. The information gathered for the following study solely reflects the individual

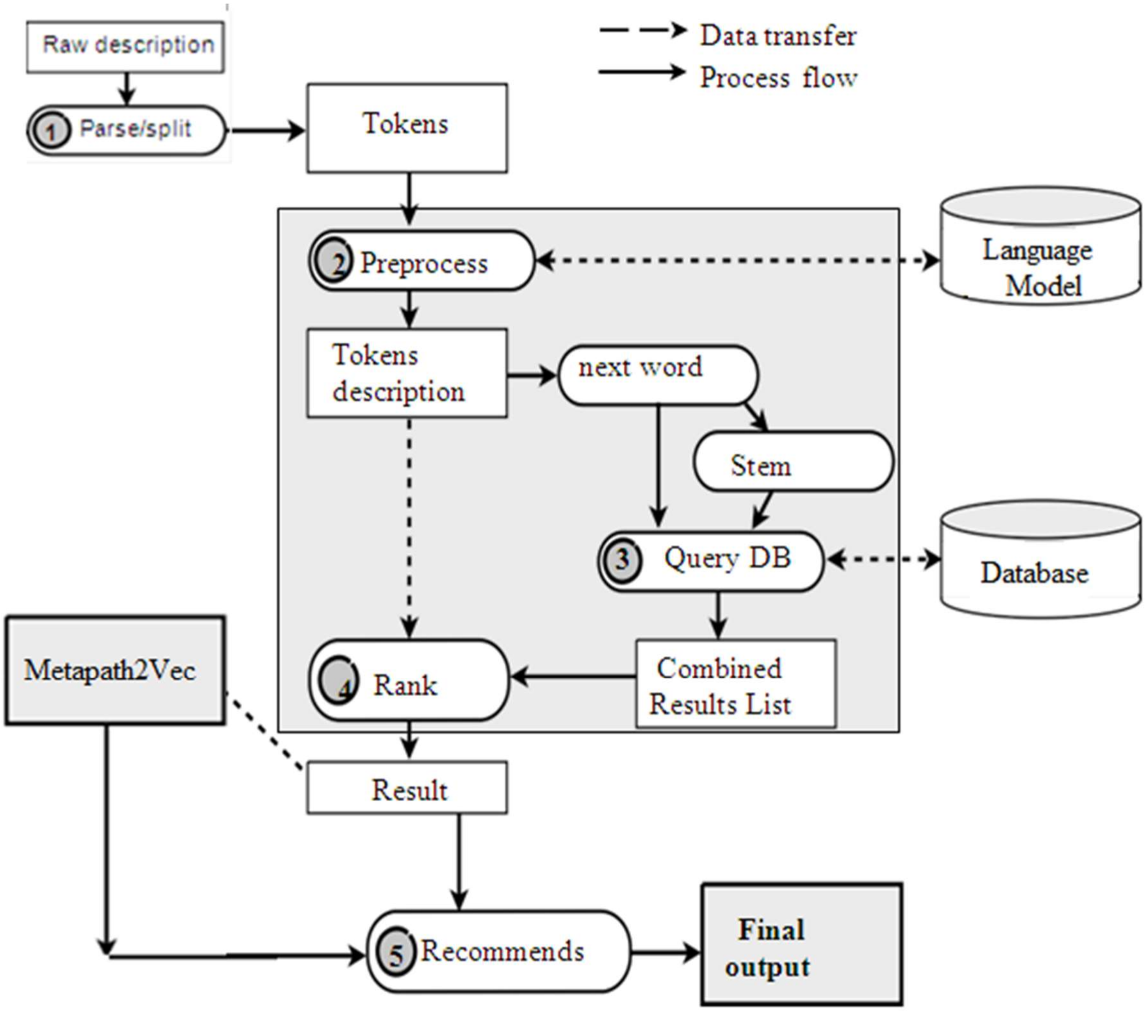

Figure 5: Process cycle of methodology and validation

experiences of the observance which can have a dramatic impact on the algorithm's performance.

Therefore, every individual medical patient can experience the same symptom through an eclectic range of words. Despite the following challenges, all the previously used techniques performed well when the current human-perceivable data is taken into account.

6. FUTURE SCOPE

Additionally, the future of the ongoing study can include tracking more complex diseases using the semantic data collected across the globe, so that the recommendation system can help the physician who is less sensitive to symptoms or false negatives in the clinical diagnosis. Along with the same, the current work can be further extended concerning the importance of the illustrated model in the healthcare sector. Scopes of future extensions include:

- Diet recommendations depending on the patient's status and related illness.
- Identification of patient-specific drug efficacy cutting across the dimensions of geographical conditions and communities
- Analysis of spatial and temporal variability of contagious disease spread.
- Future health status prediction
- Healthcare professional recommendation etc.

The current work can be further extrapolated for a greater good by implementing it on various diseases, wherein the human-based observation might be tedious. For example, the dietary adjustments of a patient suffering from gastritis could become problematic if the patient fails to identify the precursors or the foods that activate the problem of gastritis. Rather than the patient himself identifying the causation factors, a recommendation system could be constructed mapping the relationship between the food factors and their respective influences.

Along similar lines, another such example includes the condition of asthma, which is a respiratory illness. The causation factors or the disease markers could be identified by chronologically mapping the relationship between the living conditions of the patient and the initiation of the breathlessness. Similarly, augmentation of dietary precautions along with the variability of weather patterns could provide great relief to people facing environmental stimuli.

Further importance should be given to guarantee the timely availability of essential information by verifying its quality, reliability, authenticity, and privacy issues. The health recommender system is crucial for generating results such recommending diagnoses, health insurance, clinical pathway-based treatment options, and alternative medications depending on the patient's health profile, just as people use social networks to understand their health conditions. Recent studies that aim to utilise massive amounts of medical data while merging multimodal data from various sources are reviewed in order to lessen the workload and expense in healthcare. Big data analytics with recommender systems play a significant part in the healthcare industry when it comes to making decisions about a patient's health.

REFERENCES:

[1] N. R. Council and Others, “The behavioral and social sciences: Achievements and opportunities,” 1988.

[2] S. Molaei, H. Zare, and H. Veisi, “Deep learning approach on information diffusion in heterogeneous networks,” Knowledge-Based Systems, vol. 189, p. 105153, 2020.

[3] K. Wei, X. Liang, and K. Xu, “A survey of social-aware routing protocols in delay tolerant networks: Applications, taxonomy and design-related issues,” IEEE Communications Surveys & Tutorials, vol. 16, no. 1, pp. 556–578, 2013.

[4] B. Hu, H. Wang, L. Wang, and W. Yuan, “Adverse drug reaction predictions using stacking deep heterogeneous information network embedding approach,” Molecules, vol. 23, no. 12, p. 3193, 2018.

[5] C. Ruan, Y. Wang, Y. Zhang, and Y. Yang, “Exploring regularity in traditional chinese medicine clinical data using heterogeneous weighted networks embedding,” in International Conference on Database Systems for Advanced Applications, 2019, pp. 310–313.

[6] J. Shang, T. Ma, C. Xiao, and J. Sun, “Pre-training of graph augmented transformers for medication recommendation,” arXiv preprint arXiv:1906. 00346, 2019.

[7] M. M. Li, K. Huang, and M. Zitnik, “Graph representation learning in biomedicine and healthcare,” Nature Biomedical Engineering, vol. 6, no. 12, pp. 1353–1369, 2022.

[8] H.-C. Yi, Z.-H. You, D.-S. Huang, and C. K. Kwoh, “Graph representation learning in bioinformatics: trends, methods and applications,” Briefings in Bioinformatics, vol. 23, no. 1, p. bbab340, 2022.

[9] W. Liu, L. Yin, C. Wang, F. Liu, and Z. Ni, “Multitask healthcare management recommendation system leveraging knowledge graph,” Journal of Healthcare Engineering, vol. 2021, 2021.

[10] F. Ali et al., “An intelligent healthcare monitoring framework using wearable sensors and social networking data,” Future Generation Computer Systems, vol. 114, pp. 23–43, 2021.

[11] L. Jiang and C. C. Yang, “User recommendation in healthcare social media by assessing user similarity in heterogeneous network,” Artificial intelligence in medicine, vol. 81, pp. 63–77, 2017.

[12] A. Hosseini, T. Chen, W. Wu, Y. Sun, and M. Sarrafzadeh, “Heteromed: Heterogeneous information network for medical diagnosis,” in Proceedings of the 27th ACM International Conference on Information and Knowledge Management, 2018, pp. 763–772.

[13] J. G. D. Ochoa and F. E. Mustafa, “Graph neural network modelling as a potentially effective method for predicting and analyzing procedures based on patients’ diagnoses,” Artificial Intelligence in Medicine, vol. 131, p. 102359, 2022.

[14] H. Jang, S. Lee, D. M. H. Hasan, P. M. Polgreen, S. V. Pemmaraju, and B. Adhikari, “Dynamic Healthcare Embeddings for Improving Patient Care.”

[15] X. Xu et al., “Dr. right!: Embedding-based adaptively-weighted mixture multi-classification

model for finding right doctors with healthcare experience data," in 2018 IEEE International Conference on Data Mining (ICDM), 2018, pp. 647–656.

[16] Y. Wang, W. Chen, B. Li, and R. Boots, "Learning fine-grained patient similarity with dynamic bayesian network embedded RNNs," in International Conference on Database Systems for Advanced Applications, 2019, pp. 587–603.

[17] X. Huang, J. Li, and X. Hu, "Label informed attributed network embedding," in Proceedings of the tenth ACM international conference on web search and data mining, 2017, pp. 731–739.

[18] Y. Ma, S. Wang, Z. Ren, D. Yin, and J. Tang, "Preserving local and global information for network embedding," arXiv preprint arXiv:1710. 07266, 2017.

[19] W. Yu, W. Cheng, C. C. Aggarwal, K. Zhang, H. Chen, and W. Wang, "Netwalk: A flexible deep embedding approach for anomaly detection in dynamic networks," in Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining, 2018, pp. 2672–2681.

[20] A. Srivastava, S. Jain, R. Miranda, S. Patil, S. Pandya, and K. Kotecha, "Deep learning based respiratory sound analysis for detection of chronic obstructive pulmonary disease," PeerJ Computer Science, vol. 7, p. e369, 2021.

[21] R. Sadre, B. Sundaram, S. Majumdar, and D. Ushizima, "Validating deep learning inference during chest X-ray classification for COVID-19 screening," Scientific reports, vol. 11, no. 1, pp. 1–10, 2021.

[22] I. Ahmed, Z. A. Kalhoro, and Others, "Knowledge Driven Paper Recommendation Using Heterogeneous Network Embedding Method," Journal of Computer and Communications, vol. 6, no. 12, p. 157, 2018.

[23] B. Perozzi, R. Al-Rfou, and S. Skiena, "Deepwalk: Online learning of social representations," in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, 2014, pp. 701–710.