JOURNAL OF INFORMATION SCIENCE AND ENGINEERING 38, 207-221 (2022)
DOI: 10.6688/JISE.202201_38(1).0011

DIDACE: Literature Mining and Exploration of Disease-Diet Associations

RASHMEET TOOR[1] AND INDERVEER CHANA[2]
Computer Science and Engineering Department
Thapar Institute of Engineering and Technology
Patiala, 147001 India
E-mail: rtoor_phd16@thapar.edu[1]; inderveer@thapar.edu[2]

Diseases are affected and altered by different diets in multiple ways. Although diet is an important factor, there is a lack of reliable information related to disease and diet associations. The associations can only be known by reading biomedical research papers as no such dataset is readily available. Manual extraction of such associations is a time-consuming process, so in this paper, we have developed Disease Diet Associations Curator and Explorer (DIDACE) for automatically curating and further exploring disease-diet association database. A two-phase approach has been followed which includes curation of medical literature in the first phase so as to quantify the strength of association of different diseases and diets. In the second phase, generated database is further analyzed to predict the nature (harmful or helpful) of unknown associations. This is done by performing sentiment analysis and machine learning using curated database. The database, thus generated, comprises both nature and strength of Disease-Diet associations. Such databases might prove to be a useful resource for medical and health informatics researchers for understanding complex interdependencies of different foods and diseases.

Keywords: health informatics, disease-diet associations, database curation, machine learning, data acquisition and analysis

1. INTRODUCTION

Food is an essential component of disease progression as well as regression. Traditionally, home-made remedies are used commonly as cure of certain diseases. These remedies mainly comprise of food items for example, a mix of lemon, honey and warm water is good for curing cough. Similarly, the main ingredients of numerous medicines are herbs and valuable edibles. Besides its beneficial properties, diets also have a role in the progression of diseases. Excessive intake of certain food items like alcohol or fast food items is dangerous for health. Moreover, certain cooking style or combination of certain food items might further prove to be unhealthy. Thus, associations of diseases and different food items are an important factor for understanding progression of diseases. Analysis of the already known relations aid in understanding future prospects of diets. For example, if a food item is known to be related to a disease and further, the disease is known to be related to another disease, then an indirect relation between the food item and another disease can be inferred. These known associations can be analysed in different dimensions using data analysis techniques in order to gain useful insights [1]. Such information is available from different resources for example, National Cancer Institute provides booklet for managing eating problems related to cancer treatment [2], but this data is unstructured and concentrates on

Received January 29, 2020; revised July 10, 2020; accepted November 17, 2020.
Communicated by Jung-Hsien Chiang.

a single disease. Data of clinical trials undertaken in different countries is available [3] and might be used to provide information for different diseases, but such data has to be read and curated manually by domain experts. Another resource [4] which is a public education project has compiled data from food encyclopaedias and books by doctors and medical practitioners. It provides information regarding the helpful properties of foods along with references, but the data is in unstructured form and cannot be reutilized for analytics. Thus, there are two major challenges which act as a deterrent for inferring useful analytical outcomes. The first challenge is the lack of availability of structured data of disease-diet associations. A solution for this would be to extract and develop a structured dataset from available sources. One good source of such information are research papers apart from other resources like dieticians, doctors *etc.* Reading such a vast number of publications is time consuming and labour intensive, which leads to the second challenge. Hence, there is a dire need to automate the process of extraction of disease-diet associations from medical literature to further utilise it for analysis.

In this study, we aim to develop a technique for automatically extracting disease-diet associations found in medical literature and further use the extracted database for inferring more refined relations. An approach-DIsease Diet Association database Curator and Explorer (DIDACE) is proposed for automating the extraction process using curation technique and further inferring valuable associations. The main contributions of this work include:

- Design and development of an automatic technique for curation of associations between different diseases and diets from medical literature.
- Development of a prediction model for predicting nature of association of a subset of disease-diet pairs using sentiment analysis and machine learning.
- Comparative analysis of the proposed approach with other State-of-the-Art methods.

The rest of the paper is organized as: Section 2 discusses the motivation behind this work by presenting the limitations in the related work. Section 3 is a detailed description of proposed approach DIDACE comprising its materials and methods used. Section 4 discusses the experimental details and performance of results obtained using the approach. Section 5 provides conclusions along with future avenues.

2. MOTIVATION AND RELATED WORK

Many works have been undertaken to discover relationships between different diseases and foods/dietary patterns. Recently many review papers have been published in this regard. In [5], authors present a review for papers focussed on finding links between dietary patterns and cardiometabolic risk factors. The involved studies aimed to identify correlation by performing statistical analysis using data from tests and food frequency questionnaires. Another meta-analysis [6] covered 10 studies for discovering association between dietary index and cancer. The index was calculated from dietary recalls and food frequency questionnaires. In another review [7], different diets like vegetable based, ketogenic, zone diet, DASH *etc.* were analysed to find association between mood and diet. In [8], links between lifestyle factors and obesity occurrence were surveyed using National

Health and Nutrition Examination surveys. Other studies [9, 10] used statistical measures to find correlation between dietary factors and different diseases.

There are many limitations to the work done in this regard. Such studies are based on population of a particular age, ethnicity or a specific disease. Due to this, the analysis retrieves associations which are very specific. The correlations identified in these studies used traditional statistical analysis. Advanced techniques like text analysis and machine learning can promise improved results.

There are many works which have used advanced techniques, but in these, disease associations have been extracted based on parameters other than diet. Four methods have been used for curation or extraction of disease-based associations as identified in literature namely co-occurrence based, semantic analysis, machine learning and network analysis. The related works are summarized in Table 1. There are several limitations due to which use of a single method would not be efficient as summarized below:

- The literature may consist more papers of a popular topic than other topics. This might lead to a bias when only a co-occurrence based method is used.
- Use of only text mining to annotate diets in papers as harmful or helpful for diseases would result in inefficiency because there are many papers which do not state direct association but the associations can be inferred. For example, a major ingredient of a food item is found to be helpful for a disease, then it can be inferred that the food item is helpful.
- Many tools have been devised for mining of literature to extract different kinds of associations. Another limitation lies with the fact that these tools were designed for a specific data type, pre-defined ontologies or gold standard databases, thus they cannot be used for diet terms because it does not have a pre-defined database.

There are various other similar works [11-17] which have used a combination of multiple techniques to extract similar concepts, but do not target disease links. The only study [18] which has carried out analysis of disease and food associations focuses on developing their interaction networks. The networks have been statistically analyzed using network parameters to realize significant foods, disease complexity and similar diseases.

The motivation behind this work is to ease out the task of identifying relationships between diseases and diets. This work aims to reduce the task of manual curation as well as achieve good accuracy. Once accurate data sets are available in right format, it can enhance accuracy and efficiency of further analysis.

3. PROPOSED APPROACH: DIDACE

DIDACE is a two-phase approach proposed to design a technique for extracting disease diet associations automatically and develop a prediction model for further predicting the nature of extracted associations. In the first phase, a technique has been developed for extracting the count of medical abstracts in which both disease and diet terms occur together. This is done to quantify their strength of association. The extracted database is further refined by proposing a prediction model in the next phase. Some of the abstracts retrieved in first phase were read and classified as harmful or helpful. A harmful association refers to the diet escalating the effects of disease or might act as a cause for disease.

Table 1. Comparison of related work.

Ref	Year	Approach	Technique	Database	Association	Methodology	Limitation
[19]	2009	Co-occurrence based	Statistical	MeSH, PubMed	Disease-Gene	Co-occurrence based data curation	Manual curation
[20]	2014		Similarity scores	BioGrid, OMIM, gene ontology, HuGene	Disease-Disease	Integrated multiple databases	Used already developed databases
[21]	2017		–	Clinical trial, MeSH	Disease-Drug	Developed an improved MeSH vocabulary	Manual curation
[22]	2017		–	–	Disease-Microbe	Annotated associations and other descriptions	Manual curation
[23]	2015	Semantic Analysis	Scoring and Ranking	DO, HPO, Medline	Disease-Phenotype	Aber-OWL used for semantic mining of medical ontologies	Designed for pre-defined medical ontologies
[24]	2016		Latent Semantic Analysis	PubMed	Chemical-Chemical and Chemical-Disease	Analysed semantic patterns	Performance can be further improved
[25]	2017		–	Medline	Disease-Treatment	Automated generation of disease based concepts	Single disease vocabulary used for semantic schema
[26]	2017		–	Medline	Disease-Gene	DigSee tool used for text mining	Specifically designed for disease and genes association extraction
[27]	2016	Machine Learning	Shallow Linguistic Kernel	Clinical, PubMed	Disease-Drug	Curation combined with association significance filters the data, then semantic analysis further refines it	Data was collected from 1950-2011, updated data required
[28]	2016		_	PubMed	Gene-Phenotype	Triplet information extracted using text mining based machine learning	Only ten diseases covered, Performance can be improved
[29]	2016		_	PubMed	Disease-Mutation	Machine learning to identify associations using manually curated data	Redundancy in database, More robust approach required
[30]	2017		–	Medline	Disease-Gene	Use of BeFree text mining tool	Not completely automatic, experts require
[31]	2017		Maximum Entropy Classifier	MKH, PubMed, Phenominer	Genotype-Phenotype	Semi-automated approach for self-training	Better performance and enlarged training set needed
[32]	2018		Ensemble Support Vector Machine	Gold Standard Database	Disease-Gene	Semantic analysis	Not good for complex sentences
[33]	2015	Network Analysis	Random Walk	OMIM records, Ensembl	Disease-Gene	BioMart tool for querying integrated data	–
[34]	2016		Random Walk	Wikipedia, DO, MeSH	Chemical-Disease	LeadMine tool used for text mining	Better performance can be achieved

A helpful association refers to diet having soothing effects against a disease or might be helpful for its prevention. This data was further trained and used to predict harmful and helpful associations of diseases and diets in other abstracts. The various phases of DIDACE have been depicted in Fig. 1.

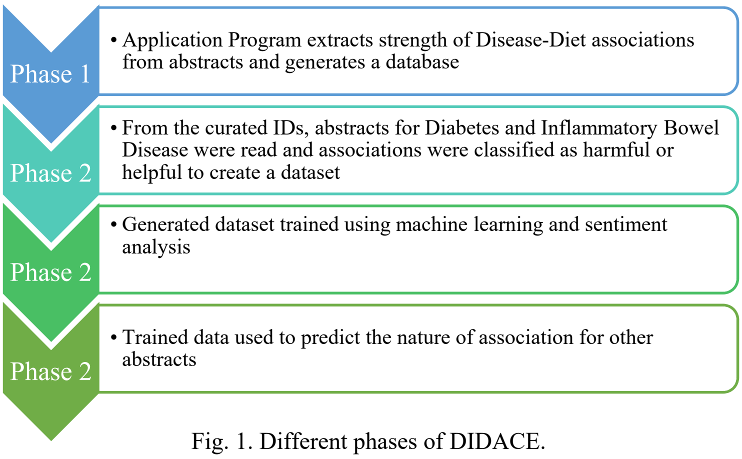

Phase 1
• Application Program extracts strength of Disease-Diet associations from abstracts and generates a database
Phase 2
• From the curated IDs, abstracts for Diabetes and Inflammatory Bowel Disease were read and associations were classified as harmful or helpful to create a dataset
Phase 2
• Generated dataset trained using machine learning and sentiment analysis
Phase 2
• Trained data used to predict the nature of association for other abstracts

Fig. 1. Different phases of DIDACE.

3.1 Materials Used

Medical Subject Headings: (MeSH) is a vocabulary which annotates research articles by representing its main topics. A hierarchy of terms has been arranged in form of tree structure with numbered notations. This was downloaded from MeSH website of National Library of Medicine (NLM) [35]. The MeSH tree contains 16 different categories, where category C has disease headings (or descriptors) while J02 has food descriptors. These categories have been used in this study so that standard terms can be utilized for curation. Thus, 4758 disease terms (including subtypes) and 154 diet terms were taken from downloaded MeSH database and stored in csv files.

PubMed: It contains citations of millions of biomedical articles from various journals. Moreover, MeSH thesaurus has indexed articles in PubMed. Thus, PubMed literature search is most suitable for extracting relevant papers containing both the disease and diet terms.

E-utilities: Manual searching of large number of terms in such a vast literature is a tedious task. Thus, National Center for Biotechnology Information (NCBI) provides an API service named E-utilities which can be used to extract required data by posting URL queries through a software. The queries are sent as URL. For example, if one needs to search all the literature in PubMed in which both the terms Coffee and Diabetes occur together, then the following query should be posted as URL:

https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=coffee [mesh] +AND+diabetes[mesh]

The [mesh] term in URL depicts that the terms are taken from MeSH vocabulary. This returns an XML with number of research papers in which both terms occur together as count, along with PubMed ids of respective papers.

Artificial Neural Network: Artificial Neural Network (ANN) is a concept replicating a human brain for solving complex problems using distributed and parallel computations. It is increasingly being used for many different applications like pattern recognition (speech, character or human face) and optimization problems. The network contains computational units called nodes/neurons connected via weighted edges. These nodes process information collected from previous nodes using an activation function. A Multi-Layer Perceptron (MLP) is a feed-forward neural network which comprises many layers of nodes with input, output and hidden layers [36]. We used this in our study because MLP is best suited for classification prediction problems that focus on tabular data.

3.2 Methods Used

In order to evade manual search and automate the process of database extraction, a novel approach DIDACE (Disease Diet Association Curator and Explorer) has been designed and developed in this work. The proposed approach is a two phased approach as mentioned below:

Phase 1: In this phase, a technique has been developed for curating research papers in which disease and diet terms occur together. It has been designed to automatically search PubMed database, extract the count of abstracts in which disease-diet terms occur together and normalize the count so as to develop a database portraying strength of association of different diseases and diets. The algorithm used to automate the extraction process has been constructed to first select all pairs of disease-diet terms from the MeSH vocabulary consecutively, then generate query URL for each pair and post it on Eutilities server, which further returns an XML page consisting the value of count of terms. The flowchart of this proposed algorithm has been depicted in Fig. 2. The count (*C*) values thus retrieved were

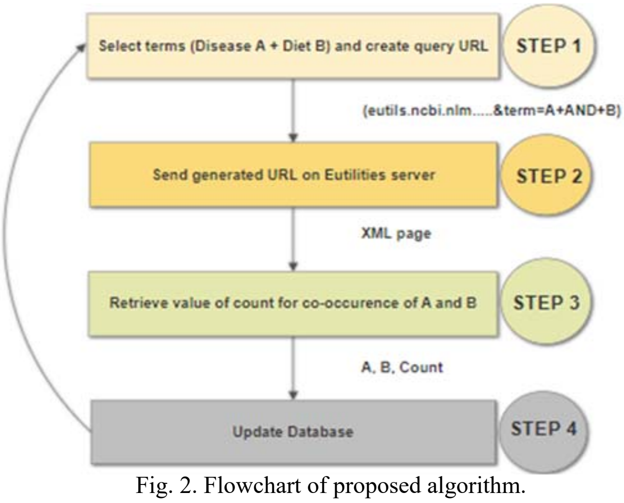

Select terms (Disease A + Diet B) and create query URL
STEP 1
(eutils.ncbi.nlm.....&term=A+AND+B)
Send generated URL on Eutilities server
STEP 2
XML page
Retrieve value of count for co-occurence of A and B
STEP 3
A, B, Count
Update Database
STEP 4

Fig. 2. Flowchart of proposed algorithm.

normalized using a formula based on Term Frequency-Inverse Document Frequency (Tf-idf). It calculates the frequency of terms taking note of their significance across all papers. Thus, co-occurrence ($C_{i,j}$) of a disease term (i) and diet term (j) has been calculated using the formula:

$$C_{i,j} = C * log(t/n)$$

where t is total number of diseases,
n is number of diseases in which diet term i occurs

Phase 2: In the second phase, a subset of abstracts pertaining to data extracted in the first phase have been used to predict the nature of association of disease-diet pairs. A bag-of-words representation was developed so that this data could be used for sentiment analysis. The steps of this phase are as follows:

1. Loading: Abstracts were selected using XML IDs and were read to classify them as harmful or helpful. These abstracts were loaded in Python and each abstract was cleaned for developing a vocabulary of tokens.
2. Cleaning: Cleaning of abstracts involved removing punctuation, numerals and known stopwords from the documents. Tokens were converted into lower case and stemming was performed. Tokens with minimum occurrence greater than 10 were taken so as to further refine the vocabulary. The most common 50 tokens with their co-occurrences are depicted in Fig. 3.

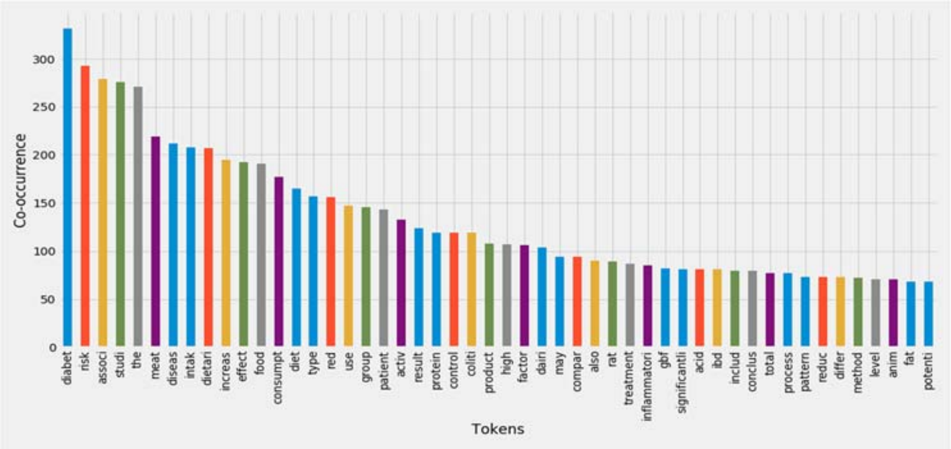


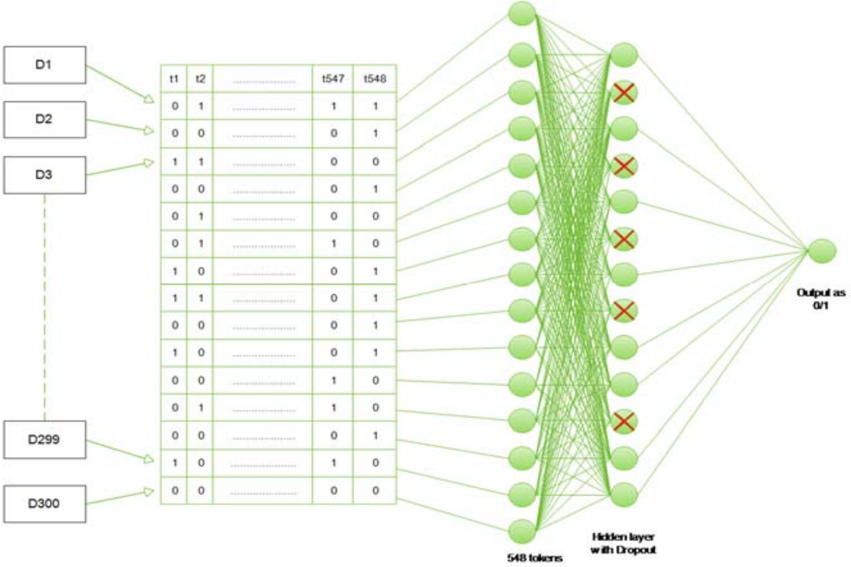
Fig. 3. Distribution of most common (50) tokens in documents.

3. Encoding: This vocabulary was further used to convert the documents into encoded vectors. The tokens in each document were scored on the basis of three vectorization methods namely, binary, count and term frequency-inverse document frequency (tf-idf). Binary method simply marks the presence (1) or absence (0) of the token whereas count method outputs the number of occurrences of each token. For example, the first document ($D1$) with n tokens can be represented as follows:

$$\begin{array}{l} t_1\ t_2\ t_3 \ldots t_n \\ D_1 = (0 \quad 1 \quad 1 \ldots \quad 0) \end{array}$$

where t_1 represents first token and a binary method of scoring is used.

4. Prediction: The encoding vectors were loaded in python for training a neural network so that prediction model can be developed. Due to tabular nature and low dimension of our data, a simple Multi Layer Perceptron (MLP) has been applied. MLP performs supervised learning for predicting class (harmful or helpful) for new abstracts. It consists of many perceptrons taking input values of features (encoded vectors in this study), which are weighted and further summed up to be used as input to an activation function. This function aids in classification decision. There can be multiple layers in an MLP, but due to low dimension of data, we chose a single hidden layer for this task. The MLP based neural network for this model is represented in Fig. 4.



t1	t2	…………………	t547	t548
0	1	…………………	1	1
0	0	…………………	0	1
1	1	…………………	0	0
0	0	…………………	0	1
0	1	…………………	0	0
0	1	…………………	1	0
1	0	…………………	0	1
1	1	…………………	0	1
0	0	…………………	0	1
1	0	…………………	0	1
0	0	…………………	1	0
0	1	…………………	1	0
0	0	…………………	0	1
1	0	…………………	1	0
0	0	…………………	0	0

Fig. 4. MLP based neural network architecture.

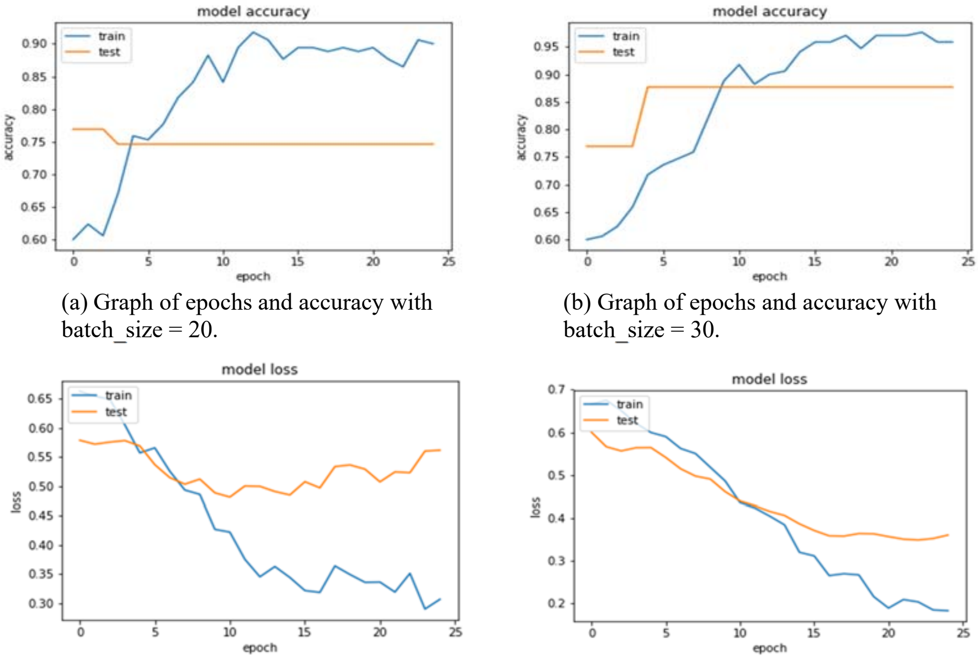
4. EXPERIMENTAL EVALUATION

4.1 Experimental Test bed Details

Phase 1: The proposed algorithm was executed as a java program. Around 3.5 lakh records were extracted, but data was filtered so as to remove records containing only 0's as correlation or other repetitions. Since this data has been curated using a program, more data could be collected in less time unlike the technique followed in [18].

Phase 2: Using the ids from XML in Phase 1, 300 abstracts for Diabetes and Inflammatory Bowel Disease were searched and read. 100 harmful and 200 helpful abstracts were identified. The cleaning, loading and encoding of abstracts for creating a vocabulary have been performed using Keras API in Python. 548 tokens were taken in vocabulary after cleaning, which were used to convert abstracts into encoded vectors. The ratio of instances used for training and testing are shown in Table 3. Since the negative samples (harmful abstracts) were less than the positive samples (helpful abstracts), we chose 70:30 ratio for splitting negative samples whereas 50:50 for positive samples. The run was repeated 10 times picking random samples for splitting in each turn so that average values can be considered. MLP with one hidden layer was trained using sigmoid activation function as the model achieves better accuracy with this function. The output layer consists of one neuron with sigmoid activation function. Adam optimizer along with binary cross entropy loss function were chosen for training.

Values of accuracy and loss were compared with different epochs and batch sizes so as to tune the parameters. Fig. 5 depicts the comparison of (i) Epochs and Accuracy, along with (ii) Epochs and Model loss, with different batch sizes. As can be seen in the figure, better accuracy has been achieved with a batch size of 30. Moreover, value of loss decreases more when this batch size is used. It reaches a minimum when the epoch value is between 20 and 25. We also introduced dropout in our MLP in order to randomly set nodes as 0 in the hidden layer. This helps in randomly selecting nodes, thus avoiding over-fitting of data. The dropout rate considered in our model is 0.2. The model also ensures that it does not suffer from exploding/vanishing gradients problems because it has only one single layer and moreover there are no large changes in loss on each update. The various hyperparameters set for this model after tuning are depicted in Table 2. The predictions were performed for new abstracts as 0 for harmful and 1 for helpful class.



(a) Graph of epochs and accuracy with batch_size = 20.

(b) Graph of epochs and accuracy with batch_size = 30.

(c) Graph of epochs and loss with batch_size = 20.

(d) Graph of epochs and loss with batch_size = 30.

Fig. 5. Comparison of different hyperparameters for tuning.

Table 2. Parameters tuned for training MLP.

Parameter	Neurons	Dropout Rate	Learning Rate	Epoch	Repetitions
Value	20	0.2	0.01	23	10

Table 3. Number of instances used for MLP training.

No. of Instances	Negative	Positive
Training	70	100
Testing	30	100

4.2 Output Evaluation

Phase 1: The final database extracted contains a total of 2,74,131 records containing 1917 different diseases and 143 diet terms. Some distinct associations were realized from this database as shown in Table 4. References for validation of extracted relations are also mentioned in the table.

Phase 2: For prediction task, 73 new PubMed abstracts pertaining to Cardiovascular and Inflammatory Bowel Diseases were chosen. Due to this, the dataset constitutes of two different diseases, thus achieving the aim of predicting for varied diseases and diets. The model predicts 1 for harmful and 0 for helpful associations.

The validation of results is performed using accuracy. The algorithm achieved different accuracies when different vectorization methods (namely binary, count and tfidf) were used for encoding as shown in Table 5. A boxplot for 3 types of vectorization methods has also been presented in Figure 6. The boxplot depicts minimum to maximum accuracies achieved for different vectorization methods. The best accuracy of 85% is achieved when tfidf is used. Another method of validation used in this study involves measure of precision and recall. Precision is a measure used to depict the accuracy of predicted positives. A false positive in our research indicates an association which is predicted to be helpful although it is not helpful. Due to this, a false positive is quite unfavorable for our research. Less number of false positives implies high precision, which in turn indicates a better model. The confusion matrix for this model has been depicted in Table 6 which is used to evaluate precision and recall. The model has a good performance as it achieves precision 88.7%, recall 81% and F1 score 84.7% as shown in Table 7. It is also important to look at these measures to identify class imbalance problem which might have occurred in our dataset due to different number of positive and negative samples.

Among these predicted associations, some test cases were taken to be validated in real life as shown in Table 8. Red meat is one such test case found to be a harmful component for cardiovascular diseases. A blog by National Institutes of Health (NIH) confirms that daily consumption of red meat triples a chemical related to heart diseases [40]. Apart from this, some helpful associations have also been predicted. Yoghurt and soy are found to be beneficial in case of Inflammatory Bowel Disease. Center for Applied Nutrition (CAN) of University of Massachusetts Medical School provides various recipes and dietary recommendations for IBD, and it recommends yoghurt and soy products for the same [41].

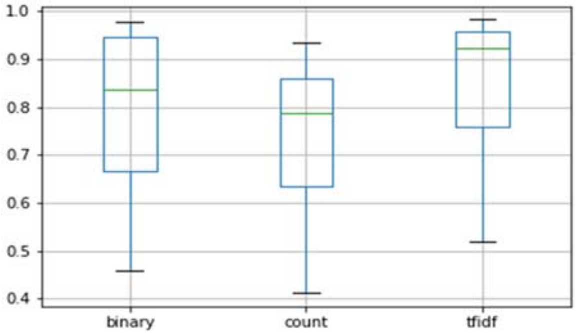


Fig. 6. Boxplot for accuracies in 3 different vectorization method.

Table 4. Sample associations extracted in Phase 1.

Diet	Disease	Reference
Avocado	Joint disease	[37]
Tea	Arthiritis	[38]
Tea	Liver disease	[39]

Table 5. Accuracies for different vectorization methods.

	binary	count	tf-idf
count	25.0	25.0	25.0
mean	0.804941	0.745412	0.857176
std	0.153381	0.150762	0.139858
min	0.458824	0.411765	0.517647
25%	0.664706	0.635294	0.758824
50%	0.835294	0.788235	0.923529
75%	0.947059	0.858824	0.958824
max	0.976471	0.935294	0.982353

Table 6. Confusion matrix for prediction model.

		Predicted	
	N=73	Negative	Positive
Actual	Negative	9	6
	Positive	11	47

Table 7. Parameters for validation of prediction model.

Parameters	Value
True Positives	47
False Positives	6
True Negatives	9
False Negatives	11
Precision	88.7%
Recall	81%
F1 score	84.7%

Table 8. Sample associations predicted in Phase 2.

Diet	Disease	Association	Reference
Red Meat	Cardiovascular disease	Harmful	[40]
Yoghurt	IBD	Helpful	[41]
Soy	IBD	Helpful	[41]

When compared with the only work [18] that has been undertaken to explore disease and food associations, it is realized that they used structural parameters of the graph generated to infer association strength. They found significance of known associations using statistical measures. Our method is different and can be considered as a next step to this work because we have used a sentiment analysis model to predict unknown associations.

5. CONCLUSION AND FUTURE DIRECTIONS

Lack of a reliable disease-diet associations database requires an effective approach for using medical literature to extract relevant data. In this study, an approach named DIDACE has been proposed to extract relationships between different foods and diseases. An automatic technique has been designed and developed for curating the association strengths of 274131 records having 1917 diseases and 143 diet terms. Further, a prediction model has been developed using machine learning and sentiment analysis to predict harmful or helpful associations from the abstracts of curations extracted earlier. The prediction model predicts harmful and helpful associations of 73 abstracts with 86% accuracy. It achieves a good performance with precision 88.7%, recall 81% and F1 score 84.7%. This method provides an efficient and time saving approach for developing disease-diet association database than a manual approach. It also provides prediction model which can be further used to extract the nature of associations of a large number of abstracts. The datasets might prove to be important assets for researchers working in the fields of computational healthcare and nutrition.

As per our knowledge, this work is first attempt to discover disease-diet associations using sentiment analysis and prediction model in literature, thus offering ample chances of improvements. It has been observed that in some cases, associations of multiple diets and diseases are present in the same research publication. In such situations, a binary classifier may not be optimal. A probabilistic classifier might outshine the ambiguities. Thus, such documents need better techniques for extracting multiple associations covering both harmful and helpful categories in the same publication. Moreover, full length papers can be used to improve the classifier further because in some cases, the associations might not be present in the abstract. Other advanced semantic analysis techniques can be utilized for better prediction accuracies.

ACKNOWLEDGEMENT

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

REFERENCES

1. M. Bhattacharyya, “Disease dietomics,” *XRDS: Crossroads*, Vol. 21, 2015, pp. 38-44.
2. National Cancer Institute, *Eating Hints: Before*, *During*, *and After Cancer Treatment*, https://www.cancer.gov/publications/patient-education/eating-hints.
3. U.S. National Library of Medicine, “ClinicalTrials.gov,” https://clinicaltrials.gov/.
4. “Healing foods reference database,” https://www.healingfoodreference.com/.
5. N. P. Rocha, L. C. Milagres, G. Z. Longo, A. Q. Ribeiro, and J. F. de Novaes, “Association between dietary pattern and cardiometabolic risk in children and adolescents: a systematic review,” *Jornal de Pediatria* (*Versão em Português*), Vol. 93, 2017, pp. 214-222.
6. N. Namazi, B. Larijani, and L. Azadbakht, “Association between the dietary inflammatory index and the incidence of cancer: a systematic review and meta-analysis of prospective studies,” *Public Health*, Vol. 164, 2018, pp. 148-156.

7. A. Arab, S. Mehrabani, S. Moradi, and R. Amani, “The association between diet and mood: A systematic review of current literature,” *Psychiatry Research*, Vol. 271, 2019, pp. 428-437.
8. S. Z. Sun and M. W. Empie, “Lack of findings for the association between obesity risk and usual sugar-sweetened beverage consumption in adults – A primary analysis of databases of CSFII-1989-1991, CSFII-1994-1998, NHANES III, and combined NHANES 1999-2002,” *Food and Chemical Toxicology*, Vol. 45, 2007, pp. 1523-1536.
9. T. Mazzeo, *et al.*, “Evaluation of a modified Italian European prospective investigation into cancer and nutrition food frequency questionnaire for individuals with celiac sisease,” *Journal of the Academy of Nutrition and Dietetics*, Vol. 116, 2016, pp. 1810-1816.
10. R. D. Mendonça, *et al.*, “Total polyphenol intake, polyphenol subtypes and incidence of cardiovascular disease: The SUN cohort study,” *Nutrition, Metabolism and Cardiovascular Diseases*, Vol. 29, 2019, pp. 69-78.
11. R. Xu and Q. Q. Wang, “Large-scale automatic extraction of side effects associated with targeted anticancer drugs from full-text oncological articles,” *Journal of Biomedical Informatics*, Vol. 55, 2015, pp. 64-72.
12. R. Xu and Q. Q. Wang, “Combining automatic table classification and relationship extraction in extracting anticancer drug-side effect pairs from full-text articles,” *Journal of Biomedical Informatics*, Vol. 53, 2015, pp. 128-135.
13. R. Xu and Q. Q. Wang, “Automatic construction of a large-scale and accurate drug-side-effect association knowledge base from biomedical literature,” *Journal of Biomedical Informatics*, Vol. 51, 2014, pp. 191-199.
14. R. Xu and Q. Q. Wang, “Toward creation of a cancer drug toxicity knowledge base: Automatically extracting cancer drug-side effect relationships from the literature,” *Journal of the American Medical Informatics Association*, Vol. 21, 2014, pp. 90-96.
15. R. Xu, L. Li, and Q. Q. Wang, “DRiskKB: A large-scale disease-disease risk relationship knowledge base constructed from biomedical text,” *BMC Bioinformatics*, Vol. 15, 2014, pp. 1-13.
16. R. Xu, L. Li, and Q. Wang, “Towards building a disease-phenotype knowledge base: Extracting disease-manifestation relationship from literature,” *Bioinformatics*, Vol. 29, 2013, pp. 2186-2194.
17. N. K. Rakhi, R. Tuwani, J. Mukherjee, and G. Bagler, “Data-driven analysis of biomedical literature suggests broad-spectrum benefits of culinary herbs and spices,” *PLoS ONE*, Vol. 13, 2018, pp. e0198030.
18. M. Bhattacharyya, S. Maity, and S. Bandyopadhyay, “Exploring the missing links between dietary habits and diseases,” *IEEE Transactions on Nanobioscience*, Vol. 16, 2017, pp. 226-238.
19. Y. Li and P. Agarwal, “A pathway-based view of human diseases and disease relationships,” *PLoS ONE*, Vol. 4, 2009, e4346.
20. K. Sun, J. P. Gonçalves, C. Larminie, and N. Pržulj, “Predicting disease associations via biological network analysis,” *BMC Bioinformatics*, Vol. 15, 2014, pp. 1-13.
21. B. Haslam and L. Perez-Breva, “Learning disease relationships from clinical drug trials,” *Journal of the American Medical Informatics Association*, Vol. 24, 2017, pp. 13-23.

22. W. Ma *et al.*, "An analysis of human microbe-disease associations," *Briefings in Bioinformatics*, Vol. 24, 2017, pp. 13-23.
23. R. Hoehndorf, P. N. Schofield, and G. V. Gkoutos, "Analysis of the human diseasome using phenotype similarity between common, genetic, and infectious diseases," *Scientific Reports*, Vol. 5, 2015, pp. 1-14.
24. C. C. Huang and Z. Lu, "Discovering biomedical semantic relations in PubMed queries for information retrieval and database curation," *Database*, Vol. 2016, 2016, No. baw025.
25. L. Wang, G. Del Fiol, B. E. Bray, and P. J. Haug, "Generating disease-pertinent treatment vocabularies from MEDLINE citations," *Journal of Biomedical Informatics*, Vol. 65, 2017, pp. 46-57.
26. J. Kim, J. J. Kim, and H. Lee, "An analysis of disease-gene relationship from medline abstracts by DigSee," *Scientific Reports*, Vol. 7, 2017, pp. 1-13.
27. D. Jang, S. Lee, J. Lee, K. Kim, and D. Lee, "Inferring new drug indications using the complementarity between clinical disease signatures and drug effects," *Journal of Biomedical Informatics*, Vol. 59, 2016, pp. 248-257.
28. A. Singhal, M. Simmons, and Z. Lu, "Text mining genotype-phenotype relationships from biomedical literature for database curation and precision medicine," *PLoS Computational Biology*, Vol. 12, 2016, pp. e1005017.
29. A. Singhal, M. Simmons, and Z. Lu, "Text mining for precision medicine: Automating disease-mutation relationship extraction from biomedical literature," *Journal of the American Medical Informatics Association*, Vol. 23, 2016, pp. 766-772.
30. A. Gutiérrez-Sacristán *et al.*, "Text mining and expert curation to develop a database on psychiatric diseases and their genes," in *Proceedings CEUR Workshop*, vol. 2017, 2017, No. bax043.
31. M. Khordad and R. E. Mercer, "Identifying genotype-phenotype relationships in biomedical text," *Journal of Biomedical Semantics*, Vol. 8, 2017, pp. 1-16.
32. B. Bhasuran and J. Natarajan, "Automatic extraction of gene-disease associations from literature using joint ensemble learning," *PLoS ONE*, Vol. 13, 2018, pp. e0200699.
33. R. Jiang, "Walking on multiple disease-gene networks to prioritize candidate genes," *Journal of Molecular Cell Biology*, Vol. 7, 2015, pp. 214-230.
34. D. M. Lowe, N. M. O'Boyle, and R. A. Sayle, "Efficient chemical-disease identification and relationship extraction using Wikipedia to improve recall," *Database*, Vol. 2016, 2016, No. baw039.
35. "Medical subject headings," https://www.nlm.nih.gov/mesh/.
36. S. Haykin, *Neural Networks and Learning Machines*, 3rd ed., Pearson, Canada, 2009.
37. A. S. A. Al-Afify, G. El-Akabawy, N. M. El-Sherif, F. E. N. A. El-Safty, and M. M. El-Habiby, "Avocado soybean unsaponifiables ameliorates cartilage and subchondral bone degeneration in mono-iodoacetate-induced knee osteoarthritis in rats," *Tissue and Cell*, Vol. 52, 2018, pp. 108-115.
38. M. Pettinger *et al.*, "Coffee and tea consumption in relation to risk of rheumatoid arthritis in the women's health initiative observational cohort," *Journal of Clinical Rheumatology*, Vol. 25, 2018, pp. 127.
39. C. Fang *et al.*, "Caffeine-stimulated muscle IL-6 mediates alleviation of non-alcoholic fatty liver disease," *Biochimica et Biophysica Acta – Molecular and Cell Biology of Lipids*, Vol. 1864, 2019, pp. 271-280.

40. National Institutes of Health (U.S.), Office of Communications and Public Liaison, NIH research matters.
41. "UMass medical school – Worcester," https://www.umassmed.edu/nutrition/ibd/gastrointestinal/ibd/.

Rashmeet Toor is currently working as a Research Scholar towards her Ph.D. in Computer Science at Thapar Institute of Engineering and Technology, Patiala. She is also presently working as Non-Tenure Lecturer at Thapar Institute of Engineering and Technology, Patiala. She completed her Masters in Software Engineering from Thapar Institute of Engineering and Technology, Patiala, in 2014. She is also a member of ACM. Her current research interests include healthcare informatics, machine learning and network analysis.

Inderveer Chana joined Computer Science and Engineering Department of Thapar Institute of Engineering and Technology, Patiala in 1997 and is currently Professor and Associate Head of the Department. She is also presently serving as Dean of Student Affairs of Thapar Institute of Engineering and Technology, Patiala. She is Ph.D. in Computer Science with specialization in Grid Computing and M.E. in Software Engineering from Thapar Institute of Engineering and Technology, Patiala and B.E. in Computer Science and Engineering. Her research interests include cloud computing, energy aware computing and software engineering. She has more than 100 research publications in reputed journals and conferences including ACM Computing Surveys, IEEE Transactions on Cloud Computing, FGCS *etc.* She is also the reviewer of many books and journals and program committee member of various conferences of repute. She has also worked on major research projects sponsored by government funding agencies like UGC, AICTE, CSIR, DST *etc.*