Automated Learning Approach for Genetic Diseases

Loay Alajramy, Adel Taweel, Radi Jarrar, Elyes Lamine, Imen Megdiche

Automated Learning Approach for Genetic Diseases

Loay Alajramy
Department of Computer Science
Birzeit University
Birzeit, Palestine
lalajramy@birzeit.edu

Adel Taweel
Department of Computer Science
Birzeit University
Birzeit, Palestine
ataweel@birzeit.edu

Radi Jarrar
Department of Computer Science
Birzeit University
Birzeit, Palestine
rjarrar@birzeit.edu

Elyes Lamine
INU Champollion, Castres, France
Toulouse University, IMT, France
Elyes.lamine@univ-jfc.fr

Imen Megdiche
INU Champollion, Castres, France
IRIT, Toulouse, France
imen.megdiche@univ-jfc.fr

Abstract—*Finding the exact gene mutations that cause a genetic disease has been a challenging task. Despite the development in information technology, the task of extracting gene-disease associations has been mainly a manual process. This is a time-consuming process, in which experts extract gene-disease associations from relevant research papers from the literature manually. The main aim of this paper is to develop an automated approach for extracting and classifying gene-disease associations from relevant literature research papers using both natural language processing and machine learning techniques. This paper extracted data from free-text literature research papers and built four different dataset formats to discover an optimal representation. Machine and Deep learning models (NB, KNN, SVM, NN, CNN, and LSTM) with TF-IDF were applied on the built datasets. As a result, the format of the dataset with (Positive and Negative) instances only, was found to be the best representation for extracting gene-disease associations with optimal accuracy between 74% and 91%. For the four dataset representations, Multilayer Neural Networks was able to predict all classes in most experiments with accuracy between 64% and 91%. From the initial results, this work highlights the need for additional work to improve both the performance of these models and the data extraction method to build more accurate and optimal dataset representation.*

Keywords— Gene-disease associations, Machine learning, Deep learning, NLP, text mining.

I. INTRODUCTION

Genetic disease is one of the main reasons for people's death. The Centers for Disease Control and Prevention (CDC) report shows that Genetic factors are nine of the ten leading reasons for death in America, and there are about 2 million people in the United States affected by a genetic condition that puts them at increased danger of heart attacks and cancer [1]. Currently, there is no way to count the number of rare human genetic diseases, because of the difficulty to identify new gene-disease associations and mutations of previous diseases [2].

The first publications of research papers on the human gene mutations field were in 1956 [3]. Since then, research in this field provided a better understanding of disease-gene association and information to improve health and prevent disease [4]. Nevertheless, our understanding of the field of human diseases and their associations with gene mutations is way from complete [4]. Ultimately, to fully understand, all disease-gene associations and all gene mutations causing genetic diseases must be identified.

Genome-wide association studies have become one of the main methods to identify the genetic bases of diseases. These led to finding a large number of DNA differences in the genome, to control, many organizations built databases to collect and save the variations, one of the popular databases is Online Mendelian Inheritance in Man (OMIM)[1], which collects the Human Genes and Genetic Disorders from research papers. Another popular database is the Moroccan Genetic Disease Database (MGDD) [2], which collects mutations in the Moroccan population [27]. DisGeNet[3] went further to provide a platform to integrate and standardize data about diseases and their associated genes [28]. The information in these types of databases allows specialists in this field to find the associations between genes and a given disease, and they can also find polymorphisms associated with susceptibility to a genetic disease [5]. This can provide information for mutation screening in genetic research laboratories and in medical diagnostic patient care services [5].

Data in these databases were collected from research papers databases such as PubMed, Web of Science, and Google Scholar and, in majority, manually analyzed by specialists to find the relations between gene mutation and diseases [5, 6]. However, this requires a continuous time-consuming effort, especially new research papers are increasingly published in this domain and thus this process requires automation to enable faster discovery and identification of not only single gene-disease associations but also cross-associations of genes to multiple diseases and vice versa. The latter challenge arises from the fact that genetic research studies, on gene disease investigation, normally focus or report on one or at best few associations. However, multiple and different cross gene-disease associations will appear in different research studies and thus requires overcoming the challenge of both discovering and resolving cross-associations from different studies, reported by different research papers. In certain cases, studies report different new associations; others may report negations of identified associations, conflicting with other studies.

Improving health must depend on complementary efforts using all forms of evidence-based interventions to reduce the burden of illness and death. Finding all disease-gene associations is a huge task of global importance. Our research aims to develop an approach for data mining to extract and classify gene-disease associations from, free text in, relevant research papers. To the best of our knowledge,

1 https://www.omim.org/

2 http://mgdd.pasteur.ma/

3 http://disgenet.org/

there is no published standardized dataset for gene-disease associations extracted from literature research papers, nor a defined optimal gene-disease dataset representation suitable for automated machine discoveries. In this paper, we focus on finding the optimal representation of gene-disease associations, with different datasets formats, and selecting the correct ones, and rejecting incorrect or less accurate ones, depending on the most accurate results of machine and deep learning algorithms.

II. Related work

Natural Language Processing (NLP) is a field of Artificial Intelligence (AI) that aims to gain value and understand the free text by using automated methods [7, 8]. The process of text classifications is essential, because of the massive amount of information in many fields included in free text, such as clinical reports, articles, posts, tweets, and many other types [9].

Many researchers deploy machine and deep learning algorithms in text classifications. Some of these approaches use supervised learning, i.e., training data, that contain input features and output labels, are provided to the algorithms to learn associations. This results in creating models that are able to predict the output labels for new instances [10]. Other approaches use unsupervised learning, i.e., do not need labeled training data to predict output features [11].

Farshchi et al. [12] used machine learning algorithms (e.g., KNN, SVM, Naive base, Neural Network) to classify medical news articles. In their work, simple Neural networks achieved the highest overall accuracy of 86%. But, using traditional machine learning, the best achieved accuracy was 80% using the K-Nearest Neighbor algorithm.

Another work that compared 7 machine learning algorithms for influenza detection from the free-text reports was presented by Pineda et al. [13]. Their results show that all machine learning algorithms almost obtained similar results with the best accuracy was 93% using the Naive Bayes algorithm. In Frunzaa et al. [14], to make sure it is effective to use text classifications of documents in the medical field, they used a dataset containing 47,274 abstracts, related to the health of the elderly, which was annotated by human reviewers to be included in or excluded from further screening. The model was built by using the Naive Bayes algorithm and achieved 63% precision.

Clark et al. [16], used text classification approaches to classify psychiatric evaluation reports to find the severity of mental disorders, and the best result, 77.86%, was obtained using NN algorithm [12]. Meenu et al. [17] suggested a model for classifying the text description of cancer tumor genetic mutations using deep learning algorithms, reporting that Recurrent Neural Network (RNN) algorithm achieved an accuracy of 70%, which is better than the other tested algorithms.

Convolutional Neural Network (CNN) has been used for text classification in many fields, as reported by Rios et al. [19]. They stated that there is a strong potential for a CNN algorithm use in biomedical text classification tasks. Wang et al. [20] also used CNN algorithm to classify clinical reports. Compared to traditional machine learning algorithms, their results show that the CNN outperformed traditional machine learning approaches and obtained an accuracy of 75%. Choudrie et al. [22], alternatively, used CNN to detect misinformation and obtained an accuracy of 86.7%. Similarly it was used for extracting Adverse Drug Reactions from free-text tweets and obtained Precision of 85.14% [23, 24], compared to other similar work [25].

As for genes related studies, several works have been presented. Nathan et al. [15], used traditional machine learning algorithms to leverage RNA-seq data from the transcript-level expression data. They achieved a considerably high classification accuracy in many cases. However, Yuhan et al. [18], developed a model to detect cancer-related genes from text within a literature-based dataset using the transformer method (BERT). They achieved good results with 80% logarithmic loss, 68.3% recall, and 70.5% F-measure. Similarly, Yujia et al. [21] built a text dataset that contains 3,740 titles and abstracts for research papers, published in PubMed[4], that are relevant to the danger of cancer for germline mutation carriers or the prevalence of germline genetic mutations and manually annotated as penetrance or prevalence. They built two models using Support Vector Machines (SVM) and CNN algorithms. Their models achieved similar results with accuracy near 89%.

III. Methodology

As presented earlier, the goal of this work is to find the optimal automated method to extract gene-disease association information from research papers retrieved from relevant literature. Our goal is to build models that classify input text (sentences) about genes and disease as *positive* or *negative* or *normal* (cf. section III.B) using machine and deep learning algorithms and evaluate the models on the created datasets.

A. Dataset

To our knowledge, there is no published dataset containing extracted text, from research papers, about the gene-disease associations.

To build such a dataset, we used the Moroccan Genetic Disease Database (MGDD) [27] to find published research about the gene-disease associations, which contain manually extracted associations, in addition to other open literature sources, e.g., google scholar. MGDD provides an ideal first step resource to select research papers, because it includes both the source research papers, in their original formats, and their manually extracted gene-disease associations by experts in the field. In this work, we selected 13 papers that have a clear positive association of a gene to disease and included some that deny or negate the association of a gene to disease.

To start the process of extracting the text from selected research papers, we used the popular OMIM database as a reference. But we had difficulty using it in the text extraction process, because the symbols for every gene, present in a single cell (in CSV file), need to be re-separated before it can be applied. The National Center for Biotechnology Information (NCBI) Gene database[5] has a more suitable distribution especially for building a programmable extraction tool, using Python. This dataset contains more than 1 million gene and mutation symbols, most of these genes and mutation symbols are in the OMIM dataset too.

4 https://pubmed.ncbi.nlm.nih.gov

5 https://www.ncbi.nlm.nih.gov/gene

However, before using NCBI, we deleted the unneeded features such as (tax_id, Org_name, GeneID, Aliases, OMIM, and others).

The next step was to build a python tool to extract text from the thirteen chosen research papers. The tool extracted sentences that contained any gene or mutation symbol that match with those in the NCBI dataset. This resulted in 320 sentences. Algorithm 1 shows the pseudocode for extracting sentences from research papers.

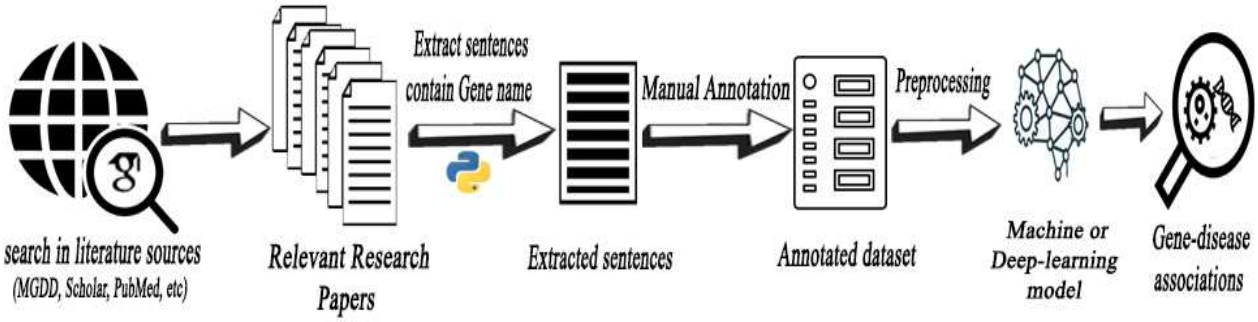

Figure 1: proposed methodology for the automated extraction of gene-disease associations.

Algorithm 1 Extracting related sentences from research papers.

1. **Start**
2. Read The CSV file (NCBI dataset) contain gene mutations symbols as **gene-symbols**
3. Initialize **Files** array contains the names of all research paper files.
4. Initialize CSV file as OutputFile.
5. Initialize variable **extracted_text**.
6. **for** all files in **Files** array:
7. Extract file text save it as **extracted_text**
8. **for** all **symbol** from **gene-symbols** in **extracted_text**
9. Search in **extracted_text** for **symbol**
10. **If** find symbol.
11. Extract the sentences (between escape characters) that contain the **symbol**.
12. Append the sentences to OutputFile.
13. End **for**
14. End **for**
15. **END**

B. *Dataset Annotation*

We manually annotated the sentences as follows:

- **Positive**: the sentence contains a gene-disease positive association, that means the gene is one of the main reasons for disease. The sentence must clearly show names of both a gene and a disease.
- **Negative**: the sentence contains a gene-disease negative association, that means the gene is not a reason for a disease. The sentence must clearly show names of both a gene and a disease.
- **Normal**: the sentence does not show any gene-disease association or do not contain names of a gene and/or a disease.

This resulted into a dataset that has 84 positives sentence, 29 negative sentence, and 207 normal sentence.

C. *Data Pre-processing*

Data pre-processing is an important step before inputting datasets into any machine learning algorithms. Data pre-processing includes data cleaning, normalization, transformation, etc.

Data cleaning entails removing empty rows and deleting none valid characters, symbols, URLs, and English stop words. All duplicate instances in the datasets were also removed. Initially, the dataset is split into 80% training and 20% test subsets taking into account that the testing subset contains data from the three classes. Table 1 shows the initial distribution for the data, named *Dataset-1*. The dataset will be modified based on the experiment results, as described in the next section.

Figure 1 summarizes the proposed methodology for the automated extraction and classification of gene-disease associations.

Table 1: Initial data distribution (Dataset-1)

	Training	Test	All
Positive	65	19	84
Negative	23	6	29
Normal	168	39	207
All	256	64	320

IV. EXPERIMENTS AND RESULTS

In this section, we describe the systematic experiments conducted to achieve the best method to automate the extracting and learning of gene-disease associations.

TF-IDF [26] is one of the most popularly used feature extracting techniques used in text datasets. TF-IDF is calculated by multiplying: 1) the number of the specific word found in a document, 2) the inverse document frequency of the word across a set of documents.

In the first experiments, we will use various machine and deep learning algorithms (NB, KNN, SVM, NN, CNN,

LSTM) with TF-IDF to investigate their ability to find the gene-disease associations on the initial dataset (Dataset-1), which was annotated with three classes (*Positive*, *Negative*, and *Normal*). Table 2 shows the results for the Dataset-1, using a carefully selected set of machine and deep learning algorithms.

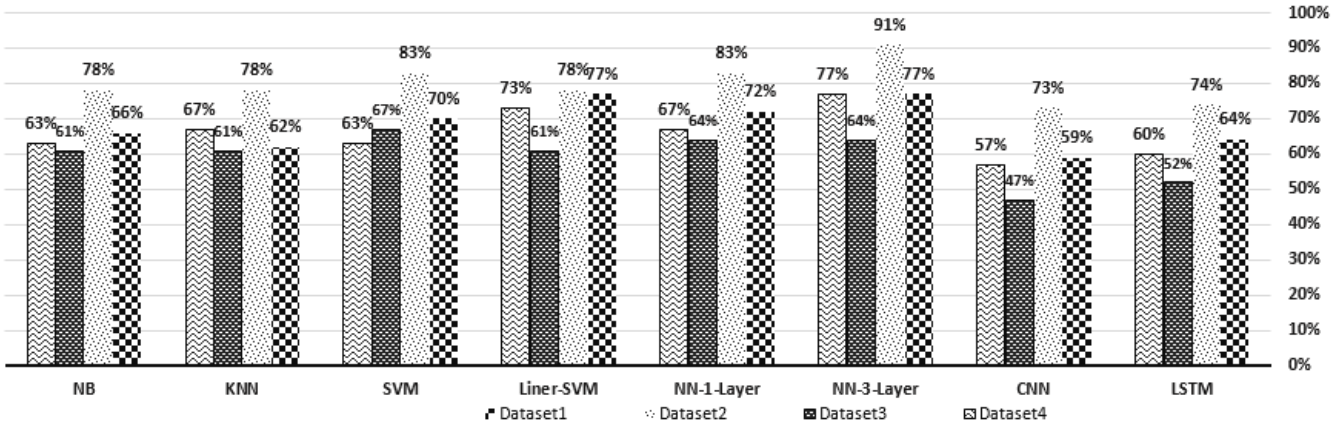

Figure 2: Accuracy for all experiments on the four dataset representations.

Table 2: Experimental results of Dataset-1

	Accuracy	Precision	Recall	F1-score
NB	66%	44%	41%	39%
KNN	62%	53%	57%	53%
SVM	70%	76%	47%	49%
Liner-SVM	77%	81%	58%	59%
NN-1-Layer	72%	69%	61%	61%
NN-3-Layers	77%	71%	61%	64%
CNN	59%	20%	33%	25%
LSTM	64%	21%	33%	26%

One of the research goals is to classify the associations between a disease and a gene, whether it is *Positive* or *Negative*, to test the suitability of the dataset format. In the second experiment, we deleted all instances annotated with the label *Normal* from Dataset-1 and repeated the experiments with the new format of the dataset, named *Dataset-2*. This dataset includes only *Positive* and *Negative* classes. Data distribution (Dataset-2) for the second experiment after deleting all *Normal* label instances is as follows: 84 *Positive*, and 29 *Negative* instances. Table 3 shows the results of the second experiment using Dataset-2.

Table 3: Experimental results of Dataset-2

	Accuracy	Precision	Recall	F1-score
NB	78%	67%	64%	65%
KNN	78%	69%	72%	70%
SVM	83%	76%	67%	70%
Liner SVM	78%	69%	72%	70%
NN-1-Layer	83%	79%	79%	79%
NN-3-Layers	91%	94%	88%	90%
CNN	73%	39%	50%	44%
LSTM	74%	37%	50%	42%

To ensure that the improved text extraction process will improve the results for extracting gene-disease associations, we built *Dataset-3* that is derived from the initial dataset (Dataset-1) by keeping the sentences that contain both name of the gene and name of the disease (both together) and deleting the sentences that do not contain both. Data distribution for Dataset-3 is as follows: 84 *Positive*, 29 *Negative*, and 51 *Normal* instances. Table 4 shows the result of the third experiment.

Table 4: Experimental results of Dataset-3

	Accuracy	Precision	Recall	F1-score
NB	61%	54%	41%	40%
KNN	61%	53%	59%	54%
SVM	67%	72%	47%	48%
Liner-SVM	61%	51%	50%	50%
NN-1-Layer	64%	58%	55%	55%
NN-3-Layers	64%	58%	52%	53%
CNN	47%	16%	33%	21%
LSTM	52%	17%	33%	23%

We built a new dataset (Dataset-4) containing 149 titles and abstracts for related scientific papers taken from the MGDD. Data distribution for Dataset-4 is as follows: 87 *Positive*, and 62 *Normal*. Table 5 shows the results of the fourth experiment.

Table 5: Experimental results of Dataset-4

	Accuracy	Precision	Recall	F1-score
NB	63%	54%	53%	51%
KNN	67%	64%	65%	64%
SVM	63%	54%	53%	51%
Liner-SVM	73%	70%	70%	70%
NN-1-Layer	67%	64%	67%	64%
NN-3-Layers	77%	77%	74%	74%
CNN	57%	28%	50%	36%
LSTM	60%	30%	50%	37%

V. DISCUSSION

As shown in experiments and results, there is superiority for the Multilayer Neural Network (NN) algorithm in comparison with the other used algorithms in most of the experiments. NN was able to predict all classes in most experiments. In the first experiments, the NN algorithm was able to predict all classes with accuracy that reached 77%, linear SVM achieved the same result too.

The results with the best accuracy were achieved in Dataset-2, in which the data included *Positive* and *Negative* classes only. In this case, the NN algorithm accuracy reached 91% and the SVM algorithm accuracy reached 83%. This led us to decrease *Normal* instances to increase the performance of the model, because of the large overlap in words between *Normal* instances and other classes. Therefore, to achieve better results, text extraction process needs to extract lesser *Normal* sentences.

Contrary to the results of other related works, it is clear that the algorithms (CNN and LSTM) were unable to classify the input data and achieved the worst results compared to other algorithms. This makes sense due to the small size of the datasets as these algorithms require large size of input training instances. When we compare the results of NN with one-layer and three-layers, we note that the increase in the number of layers in NN algorithm can lead to improved results and achieves better accuracy.

We note that the results of machine learning algorithms on Dataset-4 were disappointing. In our opinion, this is due to the long text length (number of words), where the maximum length is 163 (*mean* is 60), compared to Dataset-2, in which maximum length is 78 (*mean* is 15). Long text length usually leads to overlaps of *Positive* and *Normal* instances, which may cause low effectiveness of models. The results of Dataset-3 were also disappointing; we think this is because of the small size of the datasets compared to Dataset-1. Figure 2 summarizes the accuracy of all experiments and dataset representations.

This work highlights some of the challenges associated with extracting data from free text from research papers retrieved from the literature and creating suitable dataset representations. However, other key research challenges including automatically identifying most relevant or correct gene-disease research papers or studies from the literature, extracting and then dealing with large amount of duplicate and repeated extracted data to create a less erroneous datasets. Additionally, as more research papers get extracted, we will find the same gene associated with multiple diseases and vice-versa, multiple diseases associated with multiple genes creating more complex associations and instances across genes and diseases. Further, given that some genes and diseases will be studied more than others, this will create a natural imbalance amongst classes in the extracted data. Ultimately, a rich resultant dataset will be much more complex and will have several challenges to overcome.

The presented results are considered satisfactory as preliminary results to find the optimal representation for a dataset and find the best algorithms that can find gene-disease associations from the free text. However, as shown, we need further improvements to achieve more precise results, including obviously increasing the size of the dataset by extracting additional text instances from different published papers. The focus needs to be on achieving optimal precision, especially where the results may be used to support medical expert decisions in the field.

VI. CONCLUSION

This paper aimed to develop an automated method to extract gene-disease associations from free-text literature-derived research papers and discover an optimal dataset representation for a learning model. This was achieved by developing a free-text extraction algorithm that extracted gene-disease association sentences from scientific research papers. Then we built a dataset that contains sentences with three-classes that labelled the gene-disease associations. The dataset was manually annotated and it contains 207 *Normal*, 84 *Positive*, and 29 *Negative* instances. We built and ran experiments on four different dataset formats. We conducted experiments using various machine learning algorithms (NB, KNN, SVM, NN, CNN, and LSTM) to find the optimal dataset representation based on the best algorithm results that were able to predict the annotated classes.

We found that the Multilayer Neural Network (NN) algorithm achieved the best accuracy compared to other algorithms. For the dataset representation that contains three classes (*Normal*, *Positive*, and *Negative*), NN achieved an accuracy of 77%. The optimal dataset representation is Dataset-2 (that contains two classes: *Positive* and *Negative* only), for which NN algorithm accuracy reached 91%. Some algorithms, such as CNN and LSTM did not work well with the different tested dataset formats. Probably this is due to the small size of the datasets, which require further testing with larger datasets.

In future work, we aim to further improve the annotation process by incorporating genetic medical experts to validate the annotation of the datasets. Moreover, we aim to increase the size of the dataset by extracting additional associations from research papers. This will help in using more deep learning algorithms as they require relatively large number of training instances. We aim to improve the process of extracting text by extracting the sentences that contain both valid gene and disease names with correct associations. Additionally, transformer-based models have shown potential in other literature-based work and are worth considering on different datasets.

REFERENCES

[1] "Geography, Genetics and Leading Causes of Death | Blogs | CDC." https://blogs.cdc.gov/genomics/2014/05/15/geography/ (accessed Jun. 12, 2022).

[2] J. X. Chong et al., "The Genetic Basis of Mendelian Phenotypes: Discoveries, Challenges, and Opportunities," Am. J. Hum. Genet., vol. 97, no. 2, pp. 199–215, Aug. 2015, doi: 10.1016/J.AJHG.2015.06.009.

[3] V. M. Ingram, "A specific chemical difference between the globins of normal human and sickle-cell anaemia haemoglobin," Nature, vol. 178, no. 4537, pp. 792–794, 1956, doi: 10.1038/178792A0.

[4] T. Hartley et al., "The unsolved rare genetic disease atlas? An analysis of the unexplained phenotypic descriptions in OMIM," Am. J. Med. Genet. Part C Semin. Med. Genet., vol. 178, no. 4,

pp. 458–463, Dec. 2018, doi: 10.1002/AJMG.C.31662.

[5] H. Charoute et al., “The Moroccan Genetic Disease Database (MGDD): a database for DNA variations related to inherited disorders and disease susceptibility,” Eur. J. Hum. Genet., vol. 22, pp. 322–326, 2014, doi: 10.1038/ejhg.2013.151.

[6] J. S. Amberger, C. A. Bocchini, A. F. Scott, and A. Hamosh, “OMIM.org: leveraging knowledge across phenotype–gene relationships,” Nucleic Acids Res., vol. 47, no. D1, pp. D1038–D1043, Jan. 2019, doi: 10.1093/NAR/GKY1151.

[7] J. Hirschberg and C. D. Manning, “Advances in natural language processing,” Accessed: Jun. 26, 2022. [Online]. Available: http://science.sciencemag.org/.

[8] W. W. Yim, M. Yetisgen, W. P. Harris, and W. K. Sharon, “Natural Language Processing in Oncology: A Review,” JAMA Oncol., vol. 2, no. 6, pp. 797–804, Jun. 2016, doi: 10.1001/JAMAONCOL.2016.0213.

[9] G. Mujtaba et al., “Clinical text classification research trends: Systematic literature review and open issues,” Expert Syst. Appl., vol. 116, pp. 494–520, Feb. 2019, doi: 10.1016/J.ESWA.2018.09.034.

[10] T. Hastie, R. Tibshirani, and J. Friedman, “The Elements of Statistical Learning,” 2009, doi: 10.1007/978-0-387-84858-7.

[11] Y. Ko and J. Seo, “Automatic Text Categorization by Unsupervised Learning.”

[12] S. M. R. Farshchi and M. Yaghoobi, “Categorization of medical documents using hybrid competitive neural network with string vector, a novel approach,” Adv. Intell. Syst. Comput., vol. 180 AISC, pp. 1045–1054, 2013, doi: 10.1007/978-3-642-31656-2_144/COVER/.

[13] A. López Pineda, Y. Ye, S. Visweswaran, G. F. Cooper, M. M. Wagner, and F. Rich Tsui, “Comparison of machine learning classifiers for influenza detection from emergency department free-text reports,” J. Biomed. Inform., vol. 58, pp. 60–69, Dec. 2015, doi: 10.1016/J.JBI.2015.08.019.

[14] O. Frunza, D. Inkpen, S. Matwin, W. Klement, and P. O’blenis, “Exploiting the systematic review protocol for classification of medical abstracts,” Artif. Intell. Med., vol. 51, pp. 17–25, 2010, doi: 10.1016/j.artmed.2010.10.005.

[15] N. T. Johnson, A. Dhroso, K. J. Hughes, and D. Korkin, “Biological classification with RNA-seq data: Can alternatively spliced transcript expression enhance machine learning classifiers?,” RNA, vol. 24, no. 9, pp. 1119–1132, Sep. 2018, doi: 10.1261/RNA.062802.117.

[16] C. Clark, B. Wellner, R. Davis, J. Aberdeen, and L. Hirschman, “Automatic classification of RDoC positive valence severity with a neural network,” J. Biomed. Inform., vol. 75, pp. S120–S128, Nov. 2017, doi: 10.1016/J.JBI.2017.07.005.

[17] M. Gupta, H. Wu, S. Arora, A. Gupta, G. Chaudhary, and Q. Hua, “Gene Mutation Classification through Text Evidence Facilitating Cancer Tumour Detection,” J. Healthc. Eng., vol. 2021, 2021, doi: 10.1155/2021/8689873.

[18] Y. Su et al., “Application of BERT to Enable Gene Classification Based on Clinical Evidence,” Biomed Res. Int., vol. 2020, 2020, doi: 10.1155/2020/5491963.

[19] A. Rios and R. Kavuluru, “Convolutional Neural Networks for Biomedical Text Classification: Application in Indexing Biomedical Articles HHS Public Access,” pp. 258–267, 2015, doi: 10.1145/2808719.2808746.

[20] H. Wu and M. D. Wang, “Infer Cause of Death for Population Health Using Convolutional Neural Network,” ACM-BCB ... ... ACM Conf. Bioinformatics, Comput. Biol. Biomed. ACM Conf. Bioinformatics, Comput. Biol. Biomed., vol. 2017, p. 526, Aug. 2017, doi: 10.1145/3107411.3107447.

[21] Y. Bao et al., “Using Machine Learning and Natural Language Processing to Review and Classify the Medical Literature on Cancer Susceptibility Genes,” JCO Clin. Cancer Informatics, no. 3, pp. 1–9, Dec. 2019, doi: 10.1200/cci.19.00042.

[22] J. Choudrie, S. Banerjee, K. Kotecha, R. Walambe, H. Karende, and J. Ameta, “Machine learning techniques and older adults processing of online information and misinformation: A covid 19 study,” Comput. Human Behav., vol. 119, Jun. 2021, doi: 10.1016/J.CHB.2021.106716.

[23] A. Taweel and F. Odeh, "A Deep Learning Approach to Extracting Adverse Drug Reactions." In *2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA)*, pp. 1-6. IEEE, 2019. doi: 10.1109/AICCSA47632.2019.9035272.

[24] F. Odeh, and A. Taweel. "Semvec: Semantic features word vectors based deep learning for improved text classification." In *International Conference on Theory and Practice of Natural Computing*, pp. 449-459. Springer, Cham, 2018.

[25] G. Harsha, L. Toldo, A Rajput, J. Kors, A. Taweel, and Y Tayrouz. "Automatic detection of adverse events to predict drug label changes using text and data mining techniques." *Pharmacoepidemiology and drug safety* 22, no. 11 (2013): 1189-1194.

[26] A. R. S. Ullah, A. Das, A. Das, M. Ashad Kabir, and K. Shu, “A Survey of COVID-19 Misinformation: Datasets, Detection Techniques and Open Issues,” 2021.

[27] H. Charoute, Nahili, H., Abidi, O., Gabi, K., Rouba, H., Fakiri, M. and Barakat, A.,, “The Moroccan Genetic Disease Database (MGDD): a database for DNA variations related to inherited disorders and disease susceptibility,” Eur. J. Hum. Genet., vol. 22, pp. 322–326, 2014, doi: 10.1038/ejhg.2013.151.

[28] J. Piñero, Queralt-Rosinach, N., Bravo, A., Deu-Pons, J., Bauer-Mehren, A., Baron, M., Sanz, F. and Furlong, L.I., 2015. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, *2015*.