ROBO DOC-Machine Learning Based Disease Diagnostic Aid

Dr. Fayez Al Fayez,

Department of Computer Science and Information,
College of Science at Zulfi, Majmaah University,
Al-Majmaah 11952, Saudi Arabia

Summary

Currently genetic disorders are identified using invasive procedures which involves collecting tissue samples and analysing it but if there are complications involved in it, it would be extremely painful. Many researchers have come up with different possible solutions whilst having a few drawbacks [4,8]. Hence, a better solution has been proposed to avoid these drawbacks present in the traditional methods with better accuracy. A huge data set having multiple protein interaction and diseases caused by them would be taken as input and it will be modelled as a classification using KNN algorithm having class label as disease name, that are interacting proteins are going to be classified. Based on the symptoms, the proposed system can classify the disease and then identify the protein -protein interaction that is responsible for the disease.
Precisely, the possible treatments are identified and the best one would be suggested by the model based on its knowledge acquisition and rendition.

Keywords:

K- Nearest Neighbor(KNN), Naïve Bayes(NB), Protein- Protein Interactions(PPI), Reverse Engineering, genetic diseases.

1. Introduction

Disease diagnostic and treatment is usually a very difficult and abstract level task. That is why humans are good at it and machines fail. The major drawback of allopathic medicine is that there is symptomatic treatment but not a root cause treatment which results in temporary solution in most of the cases and not a permanent solution especially for genetic diseases. To find a better solution, identifying protein interactions that are responsible for the diseases plays a pivotal role. The diseases and protein interactions that are responsible for the cause has yielded different results by using techniques such as data mining and machine learning [9].

Data mining (DM) with classification plays a significant role in the prediction of genetic diseases and data investigation.

A huge dataset having multiple protein interactions and diseases caused by them would be taken and would be classified using classification algorithms namely KNN classification and Naïve Bayes classifier having disease name as class label [3,8]. Based on symptoms, the proposed system is capable of classifying diseases and then the protein- protein interactions that are responsible for diseases. This system can suggest treatment based on symptoms and disease identified. That is, the possible treatments are identified and the best one would be suggested by the model based on its knowledge acquisition and rendition [10,14,17].

This paper aims at diagnosing genetic diseases using symptoms and identifying protein-based interactions, responsible for them thereby identifying the root cause for a disease rather than simple symptomatic solution.

Advantages of Proposed System

- Précised diagnose system with learning capability.
- A detailed connectivity generated with proteins, disease symptoms and treatment**.**

Structure of Paper:

The rest of the paper is as organized as follows, Section 2 proceeds with literature survey and findings. The methodology is as presented in Section 3, In Section 4 the implementation details using KNN and Naïve Bayes algorithm is discussed, result outcomes are as shown in section 5. Finally, the conclusion and future work are followed in sections 6 and 7.

Manuscript received November 5, 2020.
Manuscript revised November 20, 2020.
https://doi.org/10.22937/IJCSNS.2020.20.11.3

2. Literature Survey

This section deals with in-depth knowledge about the research performed on protein interaction, methods used to achieve targets, their challenges faced, advantages and disadvantages and towards further scope for improvement.

In [1] the authors have employed a graph mining method to explore the information about PPI (protein-protein interaction) to find a proper subset of features from the samples of genes collected that help in the prediction and diagnosis of diseases, these

hub genes also assist in sample classification tasks. further, several machine learning classifiers were used in their work to evaluate various performance parameters.

A generic framework has been proposed by Yu Li et.al for disease gene prioritization, the system designed adopts training the embeddings and association prediction models in end to end basis and finds its application in computational biology.

The authors in [3] have highlighted the importance of disease and its related risks, machine learning algorithms such as KNN, CNN have been used to predict the disease from the huge amount of the medical data. As the disease prediction is a challenging task, data mining techniques has been adopted to predict the disease based on the symptoms as available from the dataset supporting early patient care.

Challenges faced due to drug use and reuse has been discussed in [4] with innovative ways that shall help to minimize development timelines and costs[14]. Several data driven and experimental approaches were also adopted towards the identification of repositioning of old drugs due to high attrition rates and to help the community.

A detailed survey was done on the aligners for the PPI networks followed by the evaluation of the biological and technological matrices in [5]. The research also leads to the importance of PPI in the field of biomedical research [14]. The scope the work was to explore the datasets that play a vital role in the performance of the aligners and help the biologists to classify the best aligner to be used.

The importance of machine learning and artificial intelligence has been discussed in [6],as the risks associated with the new diseases has become a challenging concern with many of the developed nations across the globe due to the high costs involved in health care service[10]. The authors have used several classification techniques to evaluate different performance metrics, use of genetic algorithms has helped the clinicians to classify the disease based on the data and associated feature sets.

In [7], authors have concluded in their research work the importance of promising direction for the development of new methods by using small subgraphs of the pathway from the higher-order network structures, since the computational methods rely on PPI networks with known data sets, the success is limited and the failure rate is not completely comprehended. The association of proteins with different diseases has also been addressed with disconnected pathways.

With the help of Protein Proteins Interactions (PPIs) , it is easier to understand the 3d structure of a cell and its associations. PPIs play an important role in identifying diseases in the human interaction network with related interfaces [8], the authors have made an exhaustive contribution towards different analysis of PPIs and associated results from big databases. Great success has been accomplished by implementing the deep learning algorithm [14] in sequence-based PPI prediction.

The authors in [9] have discussed the importance of machine learning in health care applications, while using large data repositories, the experts / clinicians help the patients in the early detection of risks involved and improve their survival rate [15].

Cloud-based machine intelligence has been used in [11] , with the help of smart clothing the users details are collected which make the analysis easier, the authors discuss the state of art of the terminal and cloud technology for the next generation to render the users with more reliable and efficient services.

Genetic disease identification and treatment is a complex task. Finding root cause i.e. finding protein interactions responsible for the disease is difficult and, in some cases, involves complicated painful procedures and time consuming [16,17].

3. Methodology

This section gives the insight of various UI used with the minimum system requirements towards the development of the system as discussed below:

The purpose of the designed system is to provide intelligent aid to participating physicians using machine learning techniques for identifying protein interactions that are responsible for causing genetic diseases [3,6,9,10]. The results obtained shows that the proposed model has stronger capability to predict protein interactions responsible for diseases diagnosed as compared to other

models [11]. It also contains the repository consisting of diseases, their symptoms, treatments, and proteins responsible for diseases.

The Software Module comprises with the following namely, Windows 8 / 10 OS, Python programming language, Anaconda framework with Jupyter Notebook/Spyder IDE in addition NumPy, Pandas DLL's and Flask framework to support the user interface [20-25], while the minimum hardware components required to support the functionality of the system are Core i5 processor, 1 TB HDD & 8 GB of RAM.

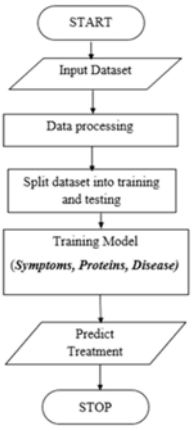


Fig. 1 System Methodology

fig 1, refers to the system methodology comprising of 4 main stages namely.

1. Data Collection

   In data collection process, the data is collected from https://snap.stanford.edu/data/ a website. Initially protein-protein interaction and Disease gene association network datasets are considered. A protein-protein interaction data consists of 21,000 proteins having 3,21,000 interactions. A Disease gene dataset consists of proteins that cause the disease. By considering disease gene dataset, symptoms-disease, disease-protein, protein-disease, and disease-treatment datasets are generated [1,2,16,17].

2. Data Pre-Processing

   In this step, data is processed. Initially null value in the dataset will be checked, if there is any null value, it will be replaced it with the mean value of that parameter and then perform the null value analysis (contains 0 as value) and replace those null values with their respective mean value. Perform a regular expression check for the data set to convert it to the correct form[7].

3. Data Splitting (split dataset into training and testing)

   In this stage, splitting of pre-processed data into train (80%) and test (20%) of data are processed as shown in fig 2.

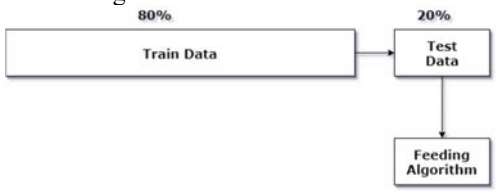


Fig. 2 Representation of data splitting

4. Constructing the training model

   This step involves "feeding" the algorithm with training data which will assist in predictive analysis, supervised and unsupervised learning are the most common styles in model training and the selection among these purely depends based on the prediction of specific attributes or group data objects by similarities.

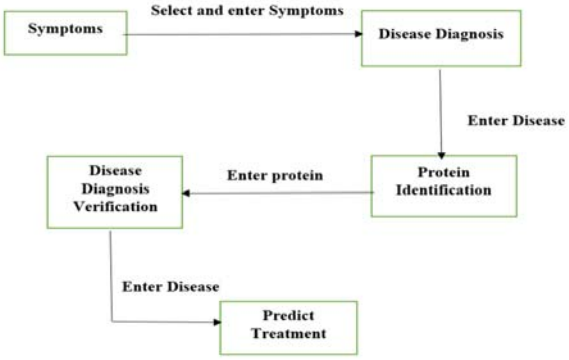


Fig. 3 Robodoc System Architecture

Fig 3 illustrates the architecture of the proposed system; Treatment Predictor is the ability to predict the treatment that has been provided to the system. For treatment prediction, naïve Bayes Classifier has been implemented[12,18].

K -Nearest Neighbor Algorithm (KNN), Artificial Neural Networks (ANN) and Naïve Bayes (NB) are used to classify the diseases and propose possible treatment using symptoms for the disease diagnosed [18,24,27]. Due to the complex nature of genetic diseases, additional care must be taken to avoid further serious complications and may also result in the death of the patient.

The proposed aid considers 20 diseases like carcinoma, Neoplasm, Heart valve disease etc. as illustrated in table 1. Results were generated using K - Nearest Neighbor Algorithm (KNN) that produced good performance in predicting the protein

interactions responsible for diseases identified using symptoms, the model achieves an accuracy of up to 100% considering 1 neighbor and 70% accuracy considering 2 neighbors. Using naïve Bayes, the model achieves 100% accuracy in predicting the disease using symptoms and predicting disease given protein, 29% in predicting treatment and 40% in predicting protein given disease.

Table 1. List of Genetic Diseases considered.

Carcinoma	Neoplasm
Squamous cell neoplasm	Ovarian diseases
Liver carcinoma	IGA
Obesity	Autistic Disorder
Kidney Failure	Brain neoplasm
Stomach neoplasm	Prostatic neoplasm
Lymphoma	Heart diseases
Schizophrenia	Peripheral neuropathy
Rheumatoid Arthritis	Adenoid Cystic Carcinoma
Salivary gland neoplasm	Diabetic Mellitus

4. Implementation

1. Data collection was a major challenge which was overcome by preparing the dataset manually. Initially protein-protein interaction and Disease gene association network dataset are taken [13]. By considering disease gene dataset manually symptoms-disease, disease-protein, protein-disease, and disease-treatment datasets are generated [15,24].

2. The challenge in KNN algorithm was solved by applying naïve Bayes algorithm[18].

3. Flask framework has used for embedding python code and HTML, CSS code to make User Interface friendlier [19,22,25].

4.1 Algorithm_K_Nearest_Neighbor :

Begin

[S 1]: Load the test and training datasets.

[S 2]: Indicate the value of K for the chosen number of the neighbors.

[S 3]: for each instance in the test data

3.1 The distance between each row of training data and test data to be computed using equation 1.

3.2 Sort them in ascending order using equation 1.

3.3. Allocate the new data points to that category for which the number of the neighbor is maximum

End

Euclidean distance : $\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$ (1)

Note: The Hamming distance as in equation 2 is used in case of the instance of categorical variables.

$$D_H = \sum_{i=1}^{k} |x_{i-y_i}| \quad (2)$$

4.2 Naïve Bayes Classifier

The Naive Bayes classifier uses the Bayes theorem of probability for prediction of unknown class. It assumes the effect of a particular feature in a class is independent of other features, finds its application especially on large data sets and supports sophisticated classification methods as illustrated in equation 3.

$$P(k\,|D\,) = \frac{P(D\,|k)P(k)}{P(D)} \quad (3)$$

Where P(k): Prior probability of k.

P(D): Probability of the data, (irrespective of the hypothesis).

P(k|D): Posterior probability i.e. the probability of hypothesis m given the data D.

P(D|k): Posterior probability of data D given that the hypothesis m was true.

4.3 Algorithm Naïve Bayes:

Begin

[S1]: compute the marginal probability for given class labels

[S2]: define the possibility of expected outcomes with each attribute for each class

[S3]: compute the conditional probability for each

class using equation 3.

[S4]: the output of prediction is the class with the highest posterior probability.

End

4.4 Gaussian naïve Bayes:

A bell-shaped curve is generated as shown in fig 4, that refers to the symmetric value of the mean features based on **Gaussian distribution.**

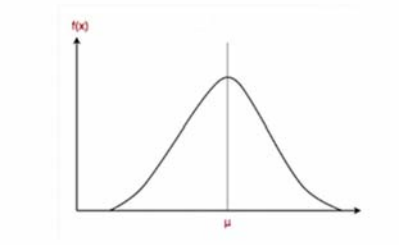


Fig. 4 Normal Distribution

Since the likelihood of the features is gaussian, the conditional probability is calculated by using the equation 4 as below.

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - y_i)^2}{2\sigma^2}\right) \quad (4)$$

5. Experimental Results

The system designed involves the Integration of four main functionalities as discussed below:

a. Protein Predication based on disease - Predicts the protein based on disease name.
b. Treatment predication based on disease- Predicts the possible treatment for disease given as input.
c. Disease predication based on proteins – Here the disease is based on multiple proteins given as input.
b. Disease predication based on symptoms-Prediction is done on different symptoms given as input.

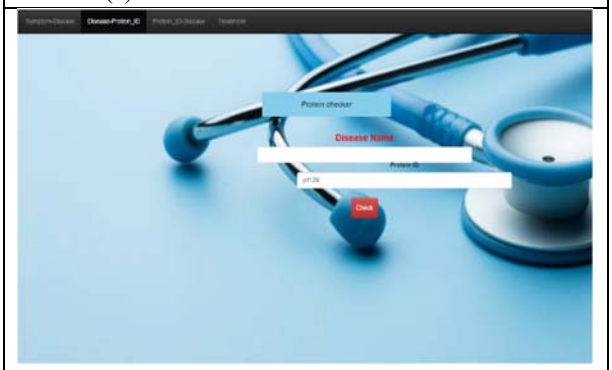
Once the integration of all the four functionalities is done, User can give input the symptoms and respective output is predicted, as listed table 2, a few testcases and the associated outputs are also verified for each type of functionality [19].

Table. 2. Test cases with associated outputs

Test Cases	Input	Expected Result	Actual Result	Status
Protein Prediction Based on Disease.	Disease name is given as input to the model.	Protein responsible for the disease is predicted.	As Expected	**PASS**
Treatment Predication Based on Disease.	Disease name is given as input to the model.	Possible treatment for the disease is predicted.	As Expected	**PASS**
Disease Predication Based on Proteins.	Multiple proteins are given as input to the model.	Based on proteins given, disease is predicted.	As Expected	**PASS**
Disease Predication based on Symptoms.	Symptoms are given as input to the model.	Based on symptoms given, disease is predicted.	As Expected	**PASS**

The following section gives the snapshots of the user interfaces used in the development of the system as shown in fig 5(a) to 5(d) [22-25].

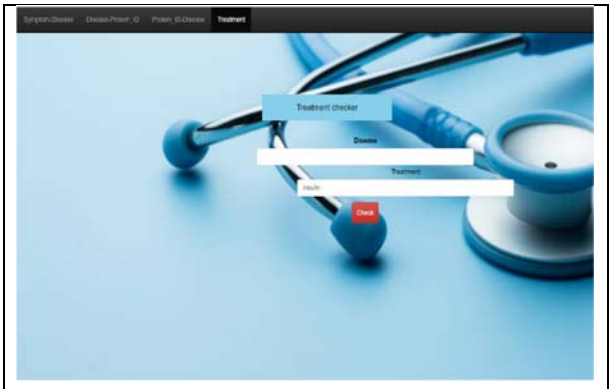
5(a). Protein Prediction based on Disease



Input : Disease name.

Output: Protein id .

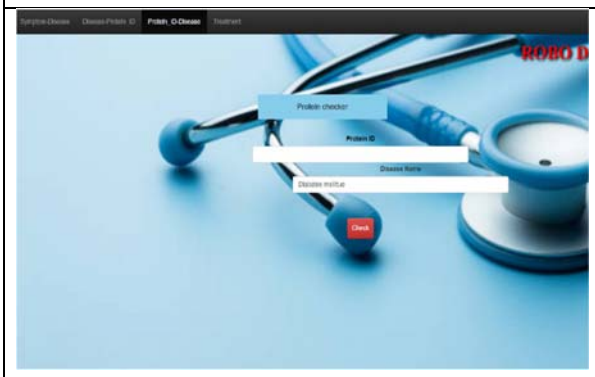
5(b). Treatment Predication based on Disease



Input : Disease name.

Output: relevant treatment is displayed.

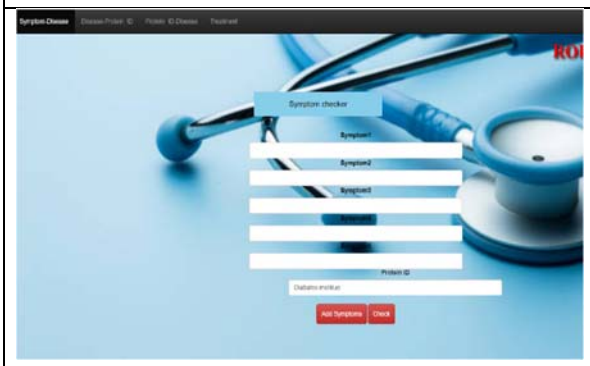
5(c). Disease Predications based on Proteins



Input : Multiple Protein id's

Output: Disease name is displayed.

5(d). Disease Predications based on Symptoms



Input : Multiple Symptoms

Output: Associated disease name is displayed.

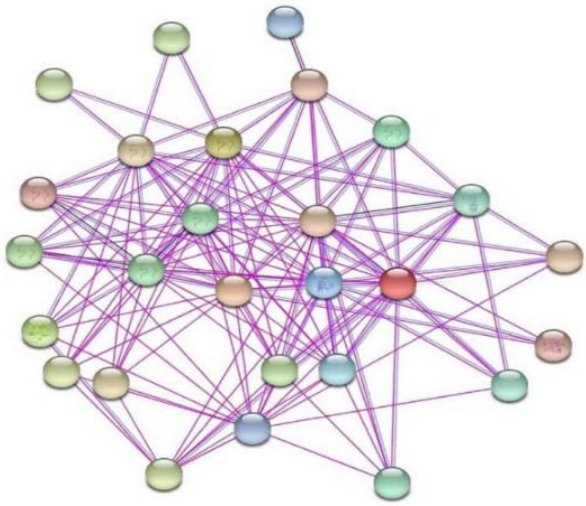


Fig. 6. Protein - Protein Interaction

Fig 6 illustrates the model of Protein - Protein Interaction, that serves a specific function and help in understanding the cell physiology, disease states and support in the drug development [1,8]. As shown in fig 7 it can be seen that the results of KNN classification algorithm yielded in 70% accuracy considering 2 neighbors and 400 proteins per disease.

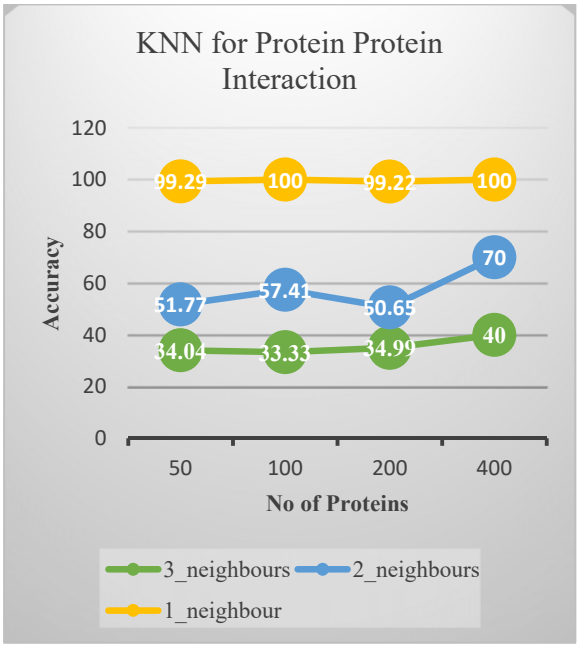


Fig. 7: Evaluation of KNN for protein - protein interaction

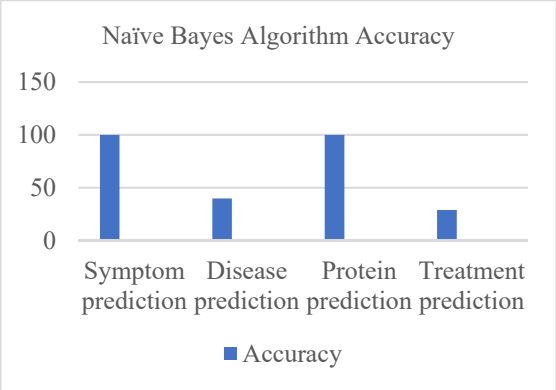


Fig. 8 Evaluation of Naïve Bayes Algorithm

The designed system is tested for all the datasets as shown in the fig 8 and the corresponding accuracies for predictions with different inputs are as obtained, **100%** for symptoms**, 29%** for treatment**, 40%** for Disease input – Protein output **and** 100% **for** Protein input – Disease output **by** applying Naïve Bayes algorithm.

6. Conclusion

Based on symptoms, the proposed system is capable of classifying diseases and the protein- protein interactions that are responsible for diseases. That is, the possible treatments are identified and the best one would be suggested by the model based on its knowledge acquisition and rendition.

The data repository makes itself a unique feature as there is no such repository that gives all the information about symptoms, proteins responsible for the diseases and the best treatment possible for genetic diseases. The model designed assists the doctors to determine the root cause of the proteins responsible for the disease and provide appropriate treatment than just symptomatic treatment.

7. Future Work

Apart from invasive procedures, diseases can be diagnosed using machine learning aid with better accuracy compared to the previous research [3,6,9,10]. For the proteins identified which are responsible for the genetic diseases, drugs can be discovered which can possibly stop the protein interactions [8,14]. Further work includes applying reverse engineering that can predict protein interactions considering only symptoms and propose possible treatment.

References

[1] P. Dutta, S. Saha and S. Gulati, "Graph-Based Hub Gene Selection Technique Using Protein Interaction Information: Application to Sample Classification," in IEEE Journal of Biomedical and Health Informatics, vol. 23, no. 6, pp. 2670-2676, Nov. 2019, doi: 10.1109/JBHI.2019.2894374.

[2] Yu Li, Hiroyuki Kuwahara, Peng Yang, Le Song and Xin Gao1, Disease gene prioritization by disease and gene embedding through graph convolutional neural networks, bioRxiv 532226; doi: https://doi.org/10.1101/532226,2019.

[3] Dhiraj Dahiwade, Gajanan Patle ,Ektaa Meshram, Designing Disease Prediction Model Using Machine Learning Approach, 3rd International Conference on Computing Methodologies and Communication (ICCMC), 2019.

[4] Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., Doig, A., Guilliams, T., Latimer, J., McNamee, C. Drug repurposing: progress, challenges, and recommendations. Nature Reviews Drug Discovery, 2019.

[5] Anooja Ali, Vishwanath R, Hulipalled, S. S. Patil, Raees Abdul kader," Alignment of Protein Interaction Networks and Disease Prediction: A Survey," International Journal of Advanced Trends in Computer Science and Engineering,2019.

[6] Satyabrata Aich, Hee-Cheol Kim, Kim younga, Kueh Lee Hui, Ahmed Abdulhakim Al-Absi, Mangal Sain, A Supervised Machine Learning Approach using Different Feature Selection Techniques on Voice Datasets for Prediction of Parkinson's Disease," 21st International Conference on Advanced Communication Technology (ICACT),2019.

[7] Monica Agrawal, Marinka Zitnik and JureLeskovec1, Large-scale analysis of disease pathways in the human interactome, Pacific Symposium on Biocomputing, 2018.

[8] Hina Umbrin, Saba Latif, "A survey on Protein - Protein Interactions (PPI) methods, databases, challenges and future directions," International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), March 2018

[9] P. S. Kohli and S. Arora, "Application of Machine Learning in Disease Prediction," 2018 4th International Conference on Computing Communication and Automation (ICCCA), Greater Noida, India, 2018, pp. 1-4, doi: 10.1109/CCAA.2018.8777449.

[10] M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," in IEEE Access, vol. 5, pp. 8869-8879, 2017, doi: 10.1109/ACCESS.2017.

[11] IM. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C. Youn, "Wearable 2.0: Enable human-cloud integration in next generation healthcare system", IEEE Commun., vol. 55, no. 1, pp. 54-61, Jan. 2017.

[12] Marinka Zitnik and Jure Leskovec, Predicting multicellular function through multi-layer tissue networks, Bioinformatics 33(14), 2017.

[13] Sun, Tanlin & Zhou, Bo & Lai, Luhua & Pei, Jianfengm," Sequence-based prediction of protein-protein interaction using a deep-learning algorithm," BMC Bioinformatics18(1), 2017.

[14] Murakami, Yoichi & Tripathi, Lokesh & Prathipati, Philip & Mizuguchi, Kenji," Network analysis and in silico prediction of protein-protein interactions with applications in drug discovery," Current opinion in structural biology,2017.

[15] Md. Tahmid Rahman Laskar, Md. Tahmid Hossain, Abu Raihan Mostofa Kamal, Nafiul Rashid, Automated Disease Prediction System (ADPS): A User Input-based Reliable Architecture for Disease Prediction, "International Journal of Computer Applications, January 2016.

[16] Antanaviciute, A. et al, "GeneTIER: prioritization of candidate disease genes using tissue-specific gene expression profiles," Bioinformatics, 2015.

[17] Ganegoda Upeksha, Wang, Jianxin , Fang-Xiang Wu , Min Li," Prediction of disease genes using tissue-specified gene-gene network,"BMC systems biology. 8 Suppl 3,2014.

[18] Feng PM, Ding H, Chen W, Lin H.," Naïve Bayes classifier with feature selection to identify phage virion proteins," Computational and Mathematical Methods in Medicine 2013.

[19] "Software testing" [Online]. Available: https://en.wikipedia.org/wiki/Software_testing

[20] "Python_(programming_language)"[Online]. Available: https://en.wikipedia.org/wiki/Python_(programming_language)

[21] "NumPy Tutorials "[Online]. Available: https://numpy.org/doc/stable/user/tutorials_index.html

[22] "Pandas" [Online]. Available: https://pandas.pydata.org/docs/reference/index.html (software)

[23] "An introduction to the Flask Python web app framework" [Online]. Available: https://opensource.com/article/18/4/flask

[24] "Python - CGI Programming" [Online]. Available: https://www.tutorialspoint.com/python/python_cgi_programming.htm

[25] "Naive Bayes and Text Classification I - Introduction and Theory",[Online]. https://paperswithcode.com/paper/naive-bayes-and-text-classification-i

[26] "flask python" [Online]. Available https://flask.palletsprojects.com/en/1.1.x/

[27] "Disease Research Papers" [Online]. Availale:https://www.papermasters.com/diseases.html