An Intelligent Framework for Feature Detection and Health Recommendation System of Diseases

Dinesh Mavaluru
d.mavaluru@seu.edu.sa
Department of Information Technology,
College of Computing and Informatics,
Saudi Electronic University, Riyadh, Saudi Arabia

Abstract

All over the world, people are affected by many chronic diseases and medical practitioners are working hard to find out the symptoms and remedies for the diseases. Many researchers focus on the feature detection of the disease and trying to get a better health recommendation system. It is necessary to detect the features automatically to provide the most relevant solution for the disease. This research gives the framework of Health Recommendation System (HRS) for identification of relevant and non-redundant features in the dataset for prediction and recommendation of diseases. This system consists of three phases such as Pre-processing, Feature Selection and Performance evaluation. It supports for handling of missing and noisy data using the proposed Imputation of missing data and noise detection based Pre-processing algorithm (IMDNDP). The selection of features from the pre-processed dataset is performed by proposed ensemble-based feature selection using an expert's knowledge (EFS-EK). It is very difficult to detect and monitor the diseases manually and also needs the expertise in the field so that process becomes time consuming. Finally, the prediction and recommendation can be done using Support Vector Machine (SVM) and rule-based approaches.

Keywords:
Feature Detection, Health Recommendation System, Expert's Knowledge, Pre-Processing, Noise Detection, Decision Support System

.1. Introduction

The Medical world is the fast undergoing revolution that seeks more information about the chronic diseases and patient's details that got affected by the diseases. Now a day, people are finding symptoms, treatment information and the diagnoses information on the internet. There is a significant increase in the number of people who search online for health and medical information. According to recent studies, 81% of U.S. adults use the Internet and 59% say they have looked online for health information regarding diseases, diagnoses and different treatments [15]. But the World Wide Web (WWW) is flooded with more irrelevant information. It is necessary to gather relevant information about the diseases and the health information of the patients.

However, information overload and irrelevant information are major obstacles for drawing conclusions on personal health status and taking adequate actions [18]. Faced with a large amount of medical information on different channels such as news sites, web forums, search engines etc. users often get lost or feel uncertain when investigating on their own. In addition, a manifold and heterogeneous medical vocabulary poses another barrier for laymen [18]. Therefore, improved personalized delivery of medical content can support users in finding relevant information. It is very hard for a patient to accurately judge how relevant the information is to their own health issues and additionally if the source of this information is reliable. One of the major challenges is to provide the quality of information which is available online.

As a solution, Health recommendation systems (HRS) supports to gather an individual's health data and to allow access to the entire users as well as for authorized medical practitioners. The responsibility of the medical practitioners is to provide reliable and relevant information as well as a solution for the patient's health issue. It is also necessary to pre-process the collected data from the user. Data pre-processing includes the expulsion of commotion and anomalies from a dataset, treatment of missing data, noisy value, redundant information and information irregularity. One of the most testing among them is to ascribe the missing data with the complete the datasets. Recuperation of these missing data vigorously influences the presentation of the data mining models. Accordingly, testing data may

Manuscript received March 5, 2021
Manuscript revised March 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.3.24

bring predisposition into the models and give incorrect results [16].

Feature selection is another significant pre-processing step in many AI applications, where it is frequently used to locate the smallest subset of features that maximally expands the exhibition of the model. Other than expanding model execution, different advantages of applying feature selection, incorporate the capacity to construct less complex and faster models utilizing just a subset of all the features by concentrating on a selected subset of features. [17].

Feature selection techniques can be divided into three categories, depending on how they interact with the classifier. Filter methods straightforwardly work on the dataset and provide a feature weighting, ranking of element as output. These methods have the advantage of being fast and independent of the classification model, but at the cost of inferior results. Wrapper methods play out a hunt in the space of features of elements, guided by the result of the model . They often report better results than filter methods, but at the price of an increased computational cost [19]. Finally, embedded methods utilize the internal information of the classification model to perform feature selection. They often provide a good trade-off between performance and computational cost [17,14].

During the past decade, the use of feature selection for knowledge discovery has become increasingly important in many domains that are characterized by a large number of features, but a small number of samples. Typical examples of such domains include text mining, computational chemistry and the bioinformatics and biomedical field, where the number of features (problem dimensionality) often exceeds the number of samples by orders of magnitude [14]. When utilizing feature selection in these domains, the robustness of the feature selection process is important. The domain experts would select a stable feature selection algorithm over an unstable algorithm to the dataset. Robust feature selection techniques would permit the domain experts to have more trust in the selected features, as in most cases these features are subsequently investigated further, requiring a lot of time and effort, especially in biomedical applications.

Expert's knowledge is implemented with the help of ontology. Ontology learning is the semantic way of knowledge representation that depends on transforming unstructured data sources into structured data. It is more related with the filed of information technology, knowledge engineering and artificial intelligence. "Ontology is a shared explicit specification of a conceptualization"[13]. In this definition, "shared" means that the information described by ontology is commonly accepted by users; "explicit" requires the precision of both concepts and their relationships clearly defined; "conceptualization" is referred to an abstract model of a phenomenon. According to the extent of dependence on the field, ontology can be subdivided into four categories, namely top level, domain, task and application ontology [9]. An Ontology defines in the basic terms and relations comprising the vocabulary of a specific area, as well as rules for combining vocabularies.

The remaining part of the paper is organized as follows, **Section 2.0** deals Literature Survey, Section **3.0** describes Framework of Health Recommendation System and **Section 4.0** gives the experimental Results. Finally **Section 5.0** concludes the paper by giving a brief glimpses into the future directions of research in this area.

.2. Literature survey

Maria Stratigi et al (2020) developed the Multidimensional Group Recommendations in the Health Domain [1]. This research work focused on giving recommendations for the group of people affected with the same kind of disease. It helps to find the semantic similarity function between users, patient medical problems, and also their education level, the health literacy, and the psycho-emotional status of the patients. Exploiting those dimensions, it provided recommendations for both most relevant and fair to groups of patients. Consequently, it is followed by a new aggregation method, accumulating preference scores. By using this method, it performed better recommendations to a small group of patients for useful information documents.

Abhaya Kumar Sahoo et al (2019) suggested the DeepReco intelligent Health Recommender System (HRS) using Restricted Boltzmann Machine (RBM)-Convolutional Neural Network (CNN) deep learning method, which provides an insight into how big data analytics can be used for the implementation of an effective health recommender engine, and illustrates an opportunity for the health care industry to transition from a traditional scenario to a more personalized paradigm in a tele-health environment [2]. The evaluation can be done through Root Square Mean Error (RSME) and Mean Absolute Error (MAE) values, the deep learning method (RBM-CNN)

which supports to reduce the errors compared to other approaches.

Jianguo Chen et al(2018) invented a disease diagnosis and treatment recommendation system based on Big Data Mining and Cloud Computing[3]. The patient's disease has been identified by Density-Peaked Clustering Analysis (DPCA) algorithm which follows clustering analysis. Apriori is used to identify Disease-Diagnosis (D-D) rules and Disease-Treatment (D-T) separately. After identifying the systems, the recommendation is given to the user by Disease Diagnosis and Treatment Recommendation System (DDTRS). The inexperienced doctors get more benefit of the system and they can recommend treatment for the patient's disease. DDTRS realizes disease-symptom clustering effectively and derives disease treatment recommendations intelligently and accurately.

Sowmya Chandrasekaran et al (2016) proposed a new pre-processing an algorithm for univariate imputation designed specifically for industrial needs[4]. Data pre-processing plays a vital role in data mining to improve the accuracy of any data. Recovery of missing data plays a vital role in avoiding inaccurate data mining decisions. We present a Seasonal and Trend decomposition using Loess (STL) based Seasonal Moving Window Algorithm, which is capable of handling patterns with the trend as well as cyclic characteristics. The performance is evaluated with a large industrial dataset and with its features. We show that the algorithm is well suited to work with various kinds of univariate datasets and is highly suitable for pre-processing of large datasets.

Meisamshabanpoor and Mehregan Mahdavi(2012) implemented a recommender system on medical recognition and treatment. This paper qualifies a medical recommender system for disease recognition and treatment [7]. The Pearson coefficient factors are used to calculate the patient's health information. The value chosen for k – the size of the neighbourhood – does not influence coverage. When the number of neighbours k taken into account is too high, too many neighbours with limited similarity bring additional "noise" into the predictions. When the value of k is too small, the quality of the predictions may be negatively affected.

Yvan Saeys et al (2008) developed a robust feature selection using ensemble feature selection techniques. They investigated the use of ensemble feature selection techniques, where multiple feature selection methods are combined to yield more robust results [12]. This classification technique has been applied and gave a better solution for high dimensional domains.

3. Framework of Health Recommendation System

The Figure 1 shows the Framework of Health Recommendation System which consists of three phases such as pre-processing, feature selection and performance evaluation. The system collects the chronicle diseases information from the user which is stored in the chronicle diseases database. The collected information is pre-processed for handling of missing and noisy data using the proposed Imputation of missing data and noise detection. The selection of features from the pre-processed dataset is performed by proposed ensemble-based feature selection using an expert's knowledge. Finally, the prediction and recommendation can be done using Support Vector Machine and rule-based approaches.

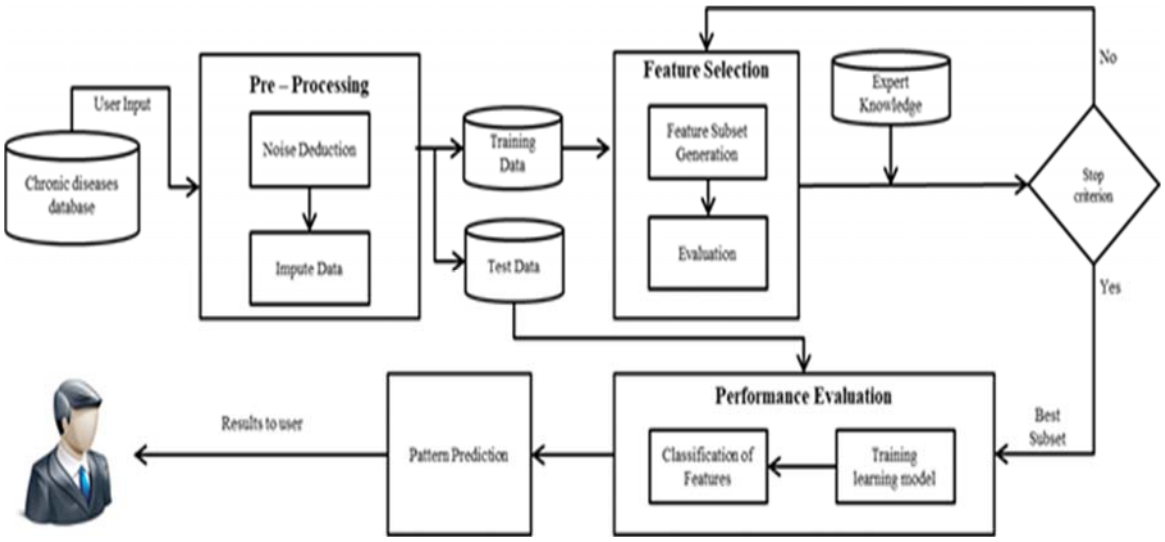

Fig.1: Framework of Health Recommendation System

3.1 Pre-Processing

Handling missing data is the unavoidable aspect of data analysis. When the user is missed to fill the required information, then it is difficult to give the most relevant information for their query. The imputation method is one of the best solutions for handling missing and noisy data. Few imputation methods are listed below.

- Likewise Deletion method removes the incomplete or empty responses from the user.
- Zero imputation is used when the data is considered as incorrect.
- Mean imputation method calculates the mean of all the within the same attribute and then imputed in the missing data cells.
- Multiple imputation method collects the information from all the variables to derive a new variable which act as an imputed value for the missing value.
- Regression imputation uses linear regression function to calculate from the values within the same attribute considered as the dependent variable and all other variables are considered as an independent variable. The calculated dependent variable is imputed for the missing value.
- Stochastic regression imputation involves a two-step process in which the relative frequency for each response is collected and the imputed value is obtained from the observed data.
- Predictive mean imputation uses a linear equation to find out the imputed value for the missing data.

In order to detect noise, this research proposed an improved noise detection technique. The noisy instances from chronic dataset must be detected and eliminated in order to improve the classification accuracy. The presence of noisy instances in data is a fundamental issue for classifying with many potential negative consequences. We employ both the naive Bayes classifier and Laplacian estimator which dynamically determine the noise threshold according to individual's history. The Laplace estimator can be written as

$$\rho(C = c_i) = \frac{n_c + k}{N + nxk} \tag{3.1}$$

Where n_c is the number of instance satisfying C = c_i, N is the number of training instances, and n is the number of classes and k = 1. Based on the Probability values, check whether instances are purely classified or misclassified and save all the pure and misclassified ones separately. Then compare noise threshold value with each instance in misclassified probability. If noise threshold value is greater than the probability value, then save the instances as noise and remove it.

3.2 Feature Selection

Feature selection is an important method of pre-processing technique in which the features of the information has been retrieved. Automatic feature extraction is the task to identify features of the attributes and their relationship that can describe the meaning of the feature. Based on the model, the features should be extracted automatically and with the minimal number of human involvements. The goal of automatic feature extraction is to apply the power and speed of computation to the problems of access and discoverability, adding value to information organization and retrieval. The aim of feature selection is to extract a set of features of the attributes.

The proposed ensemble feature selection model aims at improving the feature subset selection and to increase the accuracy of predicting features, using a specific number of attributes. Instead of relying on purely automated or purely expert-based feature selection, another approach to combine different methods is proposed. According to different rules, the learning results of multiple optimal feature subset candidate sets are aggregated to obtain the optimal feature subsets. Finally, classification algorithm with good performance is used to verify the proposed algorithm. Feature selection process through different algorithm such as Pearson Correlation Coefficient, M5 Algorithm and Infogain Algorithm.

Pearson Correlation Coefficient between variables X and Y is calculated by relevance analysis. The variable X has been taken as the first feature from the S list. Then find and remove all features for which X is approximately equivalent according to the Pearson $\chi 2$ test. Set the next remaining feature in the list as X and repeat for all remaining features in the S list. In M5 Algorithm, initially coefficients to each data point for being in the clusters must be assigned randomly. Repeat until the algorithm has converged and each time compute the centroid for each feature. And also for each data point, compute its coefficients of being in the feature subset.

In Infogain, the discernibility matrix for the selected dataset is computed using the discernibiliy function. The attribute has to be selected which belongs to the large number of conjunctive sets, numbering at least two, and apply the expansion law. Then substitute all strongly equivalent classes for their corresponding attributes and calculate the Information gain for the simplified discernibility function contained attributes. Finally, choose the highest Gain value and add it to the reduction set, and remove the attribute from the discernibility function. The fig 2 shows the Ensemble Feature Selection algorithm.

```
Algorithm: Ensemble Feature Selection

Input:
n ranking lists (list 0 to n-1) and each list has k features.
Output:
  1. An array F containing features and their rank in each ranking list,
     count, and mean rank.
  2. An ensemble list E.

Initialize E and F to empty
FOR each ranking list i
     FOR each feature in ith ranking list
          IF the feature is not in F
               Add the feature and its rank in list i to F
               FOR list j, j is from i+1 to n-1
                    IF the feature is in the list j
                         Add the rank of the feature in list j to F
                    ENDIF
               ENDFOR
          ENDIF
     ENDFOR
ENDFOR
FOR each feature in F
     Calculate frequency and mean rank of the feature
ENDFOR
Sort the features in F based on their frequency, if same frequency, sort
by mean rank;
select the top k features and assign the features to list E.
```

Fig.2: Ensemble Feature Selection Algorithm

3.3 Classification and Prediction

Classification is a data mining technique used to predict group membership for data instances. For the purpose of classification, the k-nearest neighbor widely used algorithm based on different learning principles are applied. The 10-fold cross validation is similar to the repeated holdout validation. The advantage is that all available samples are eventually used for both training and testing because of systematic data partitioning. The fig 3 shows the Process of classification and prediction.

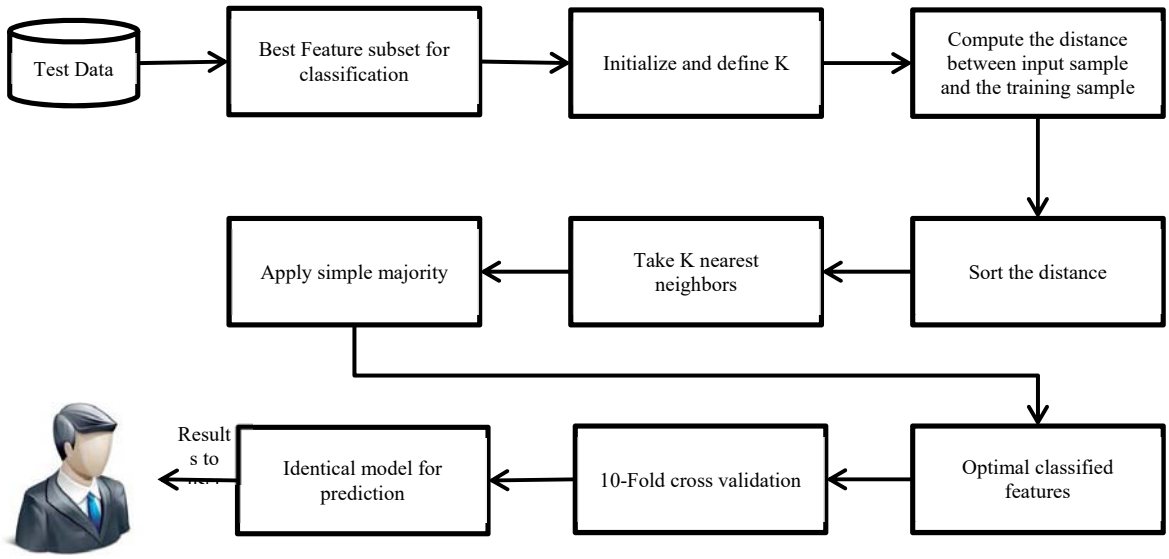

Fig 3: Process of classification and prediction

In validation, the set of features generated using the feature selection of the proposed framework and fed as inputs to the classifier. All the features are arranged in order of their coefficient of variances. This approach also requires less computational time than the other methods. The selected feature subset from feature selection method will be used to build a classifier model using test dataset, and the quality of the features are assessed by evaluating performance of the classifier.

4. Experimental Result

The quality of selected feature subset from proposed method is evaluated in a way by evaluating performance of classifiers that are built based on those features. In addition, evaluation by domain expert is also provided to give additional insight on quality of selected features and suggestions for future improvement. The step is repeated 10 times for each class and the outputs are recorded. The classifier was trained for each subset class using the five features as the input and the default value of the parameter n= 0.1. The outcome of the proposed research work is evaluated by different parameters such as accuracy, sensitivity, specificity, False Positive Rate (FPR) and False Negative Rate (FNR). The metrics are used for this parameters are described as True Positive (TP), True Negative (TN), False Positive (FP), False Negative (FN).

The table 1 shows the implementation results for the different feature selection algorithm.

Table 1: Result of feature Selection Algorithm

FS Algorithm	Accuracy	Time
FS-PCC	0.9167	0.87
FS-M5	0.9021	0.76
FS-IG	0.8934	0.81
Ensemble Model	0.9986	0.65

Various existing methods are compared with the proposed model for validating the performance results in chronical disease feature selection. Table 2 describes the values of various metrics for the proposed model with existing methodologies such as accuracy, specificity and sensitivity are computed. Table 3 shows the Mean accuracy of 10-fold cross-validation accuracies achieved on each dataset for KNN and the graphs are shown in the figure 4 and 5.

Manuscript received March 5, 2021
Manuscript revised March 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.3.24

Table 2: Performance Measure

Performance Measure	FS-PCC	FS-M5	FS-IG	Ensemble Approach
Sensitivity	0.2556	0.7654	0.8790	0.9432
Specificity	0.6287	0.4598	1.4323	0.9698
Accuracy	0.5498	0.8854	0.0956	0.9376

Table 3: accuracy of 10-fold cross-validation

KNN Classifier	FS-PCC	FS-M5	FS-IG	Ensemble Approach
Diabetes	0.56	0.59	0.64	0.71
Heart-Statlog	0.54	0.84	0.69	0.90
Ionosphere	0.57	0.75	0.78	0.86
Mammogram	0.76	0.83	0.87	0.95
Sonar	0.58	0.68	0.60	0.59

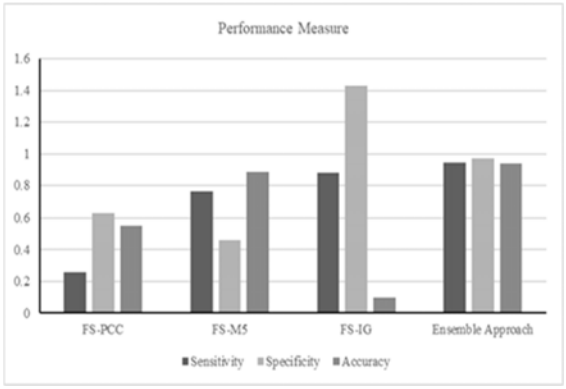

Fig 4: Performance Measure

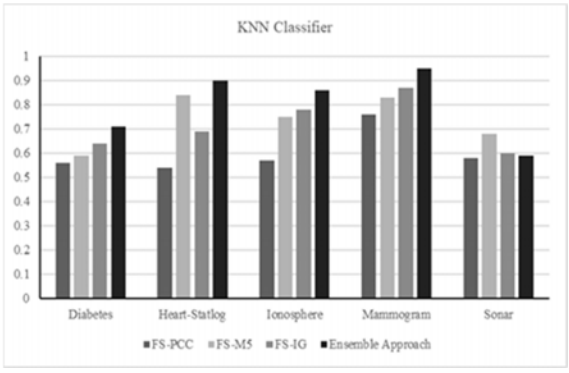

Fig 5: Comparison of KNN Classifier

5. Conclusion

Health Recommendation systems are one of the new innovating technologies for inferring advantageous information for a patient's health issue. These systems find recommended hospitals by calculating the similarity of patients' choices. Therefore, they play an important role in the medical sector In this research work, we focus on health recommendation system which deals with the missing data and noisy value in the health domain using support vector machine. For rectifying the missing value and noisy data issue, the imputation of missing data and noise detection based pre-processing algorithm is used. Feature has been extorted using ensemble-based feature selection using an expert's knowledge. The quality of selected feature subset from each method is evaluated in two ways by evaluating performance of classifiers that are built based on those features, and by evaluating their stability. Current limitations on this research project opens up many room for improvement to be considered for future research. Test the proposed method with different kind of classifier, or the same with possibly better setting. The Feature selection method focuses on stability and interpretability and Implemention technique to improve data sampling as a way to deal with data imbalance.

6. Reference

[1] Maria Stratigi , Haridimos Kondylakis and Kostas Stefanidis, "Multidimensional Group Recommendations in the Health Domain", Algorithms 2020, 13, 54; doi:10.3390/a13030054.,2020.

[2] Abhaya Kumar Sahoo , Chittaranjan Pradhan , Rabindra Kumar Barik and Harishchandra Dubey, " DeepReco: Deep Learning Based Health Recommender System Using Collaborative Filtering", Computation 2019, 7, 25; doi:10.3390/computation7020025.,2019.

[3] Jianguo Chena, Kenli Lia,b , Huigui Ronga , Kashif Bilalc , Nan Yangd and Keqin Li, "A Disease Diagnosis and Treatment Recommendation System Based on Big Data Mining and Cloud Computing", Int. Journal of Information Sciences, 2018.

[4] Sowmya Chandrasekaran, Martin Zaefferer, Steffen Moritz, Jörg Stork andThomas Bartz-Beielstein, "Data Preprocessing: A New Algorithm for Univariate Imputation Designed Specifically for Industrial Needs", Proc. 26. Workshop Computational Intelligence, Dortmund, 24.-25.11.2016.

[5] P. Kang, "Locally linear reconstruction based missing value imputation for supervised learning," Neurocomputing, vol. 118, pp. 65–78, 2013.

[6] Agarwal D., Chen B., Elango P., Ramakrishnan R. "Content recommendation on web portals", Commun. ACM. 2013; 56:92–101, 2013.
[7] Meisamshabanpoor and Mehregan Mahdavi, "Implementation of a Recommender System on Medical Recognition and Treatment", International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 2, No. 4, August 2012.
[8] Berant, Jonathan, Ido Dagan and Jacob Goldberger, "Global learning of typed entailment rules. In: Forty Ninth Annual Meeting of the Association of Computational Linguistics": Human Language Technologies, Portland, Oregon, USA, pp. 610–619, 19–24 June 2011.
[9] Roitman H., Yossi M., Yevgenia T., and Yonatan M. "Increasing Patient Safety Using Explanation-driven Personalized Content Recommendation", Proceedings of the 1st ACM International Health Informatics Symposium (IHI '10), ACM; New York, NY, USA. pp. 430–434., 11–12 November 2010.
[10] Sommerhalder K., Abraham A., Zufferey M.C., Barth J., and Abel T. "Internet information and medical consultations: Experiences from patients and physicians' perspectives" Patient Educ. Counsel.77:266–271. doi: 10.1016/j.pec.2009.03.028.,2009.
[11] Swan M., "Emerging patient-driven health care models: An examination of health social networks, consumer personalized medicine and quantified self-tracking", Int. J. Environ. Res. Public Health. ;6:492–525. doi: 10.3390/ijerph6020492,2009.
[12] Yvan Saeys, Thomas Abeel, and Yves Van de Peer, "Robust Feature Selection Using Ensemble Feature Selection Techniques " , W. Daelemans et al. (Eds.): ECML PKDD 2008, Part II, LNAI 5212, pp. 313–325, 2008. c Springer-Verlag Berlin Heidelberg 2008.
[13] Aussenac-Gilles, N., Despres, and S., Szulman, S., "The TERMINAE method and platform for ontology engineering from texts", In:Proceeding of the 2008 Conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge, pp.199–223.,2008.
[14] Saeys, Y., Inza, I., and Larra˜naga, P.: "A review of feature selection techniques in bioinformatics", Bioinformatics 23(19), 2507–2517, 2007.
[15] McMullan M., "Patients using the Internet to obtain health information: How this affects the patient-health professional relationship". Patient Educ. Couns., 63, 24–28,2006.
[16] Brown, M.L. and Kros, J.F. "The impact of missing data on data mining.", Data mining: Opportunities and challenges, 1, pp.174-198, 2003.
[17] Guyon, I., and Elisseeff, A. , "An Introduction to Variable and Feature Selection",Journal of Machine Learning Research 3, 1157–1182 ,2003
[18] Hardey M. "Doctor in the house: The Internet as a source of lay health knowledge and the challenge to expertise" , Sociol. Health Illness. ;21:820–835. doi: 10.1111/1467-9566.00185. 1999.
[19] Kohavi, R., and John, G."Wrappers for feature subset selection", Artif. Intell. 97(1-2), 273–324 ,1997.

Acknowledgements:

I would like to express our gratitude to the Saudi Electronic University in providing a platform for my research. Also, we would like to express my appreciation to all the colleagues of my department for their constant support.

Dr. Dinesh Mavaluru is currently working as Assistant Professor in the Department of Information technology at Saudi Electronic University, Saudi Arabia. He received a Ph.D from B S Abdur Rahman University. His research interests span both data science and network science. Much of his work has been on improving the understanding, design, and performance of parallel and networked computer systems, and healthcare systems mainly through the application of data mining, statistics, and performance evaluation. Committed to helping students identify and develop their own passions while becoming successful and confident scholars and learners. Exceptional track record of research success with multiple published articles in highly indexed journals and conferences.