Symptom Recommendation using Collaborative Filtering and Disease Prediction using Support Vector Machine

Akshay Kamath
KJ Somaiya College of Engineering
Mumbai, India

Amogh Parab
KJ Somaiya College of Engineering
Mumbai, India

Neeraj Kerkar
KJ Somaiya College of Engineering
Mumbai, India

ABSTRACT
Early diagnosis and identification of diseases play a vital role in the field of medicine. With the emergence of powerful machine learning techniques, it is now possible to derive greater insights from the available data. This paper discusses how one such machine learning technique can be used to recommend symptoms to users. The proposed system allows users to enter symptoms and uses machine learning techniques to recommend similar symptoms. Another machine learning technique for classification is discussed which is used predict the possibility of having a disease. The results demonstrate the effectiveness of different machine learning techniques on the given data.

General Terms
Machine Learning, Recommendation System, Collaborative filtering, Classification

Keywords
Support Vector Machine, Symptom Recommendation, Jaccard coefficient, Django, Electronic medical record, Disease prediction

1. INTRODUCTION
Recent advances in information technology have changed the way health care is carried out and documented.[6] Data is available in form of Electonic Medical Records(EMR) which can be mined for yielding useful patterns that can help in clinical research and decision making.The EMR contains useful information about a patient such as the history of hospital visits, diagnoses, demographic information etc.[4]As these data obtained may contain a large amount of redundant and irrelevant attributes or features, feature selection must be done to select a subset of features in order to build a predictive model.

Use of recommendation algorithm for recommending user to perform a specific activity that will improve user's health, based on his given health condition and set of knowledge derived from the history of the user and users like him has become popular nowadays.[3] Machine learning algorithms are being used nowadays in recommender systems for providing better recommendations. Collaborative filtering is used in recommender systems which considers user data when processing information for the recommendation.[1]Chronic diseases such as heart disease, breast cancer, and diabetes are on the rise and there is a growing need for a system that helps diagnose these diseases at an early stage.

Machine learning and neural networks have recently gained traction as reliable and robust prediction systems that help provide a deeper insight into the statistical data.Using data from the past, it is possible to design a machine learning system that helps doctors diagnose diseases with a higher level of accuracy. The proposed system makes use of collaborative filtering for suggesting symptoms to the patient and machine learning algorithms like support vector machine for disease prediction.In order to give users an interactive and easy to use interface, the web platform can be used.

The organization of this paper is as follows. Section 1 is the introduction. Section 2 describes relevant background knowledge and concepts related to disease prediction, classification, support vector machines etc. In section 3, the architecture of the proposed system has been explained in detail.Section 4 describes how disease prediction is done.Section 5 describes the datasets used for testing, the features that are taken into consideration for making predictions, and details about the training process. The results of our tests are presented in section 6. Finally, in section 7, the conclusions are presented, along with further lines of work.

2. LITERATURE SURVEY
Most existing systems for health monitoring allow patients to interact with doctors but don't perform automatic disease prediction. Those systems that do provide disease prediction do so only for a few diseases. Also, patients need to enter different kinds of data for each of these different diseases.

Use of classification algorithms has been common in disease prediction. Research paper referred to as part of the literature survey compared Support Vector Machine and Artificial Neural Network classifiers for prediction of heart coronary disease which concluded that SVM was the more viable option. [2] Some other systems have used Artificial Intelligence and neural networks for disease prediction. [5]

The system proposed in this paper not only allows patients to interact with doctors but also provides a feature where patients can enter their symptoms and the system can predict the possibility of having a disease from among a wide range of diseases such as arthritis to cancer. There is a uniform user interface which takes patient information and symptoms as input. No separate data needs to be taken for predicting different diseases. For each of these diseases, support vector machine classifiers have been trained for predicting the possibility of having that disease.The system also makes the process of entering symptoms easier for the patient, based on the symptoms that the patient has currently entered the system recommends to the patient other similar symptoms that the patient is likely to have. Collaborative filtering has been used for finding related symptoms.

3. SYSTEM ARCHITECTURE
The proposed system architecture is shown in Figure 1. The entire architecture can be divided into following parts-SVM module, symptom recommendation module, web interface

module and the Python backend containing the disease prediction module.

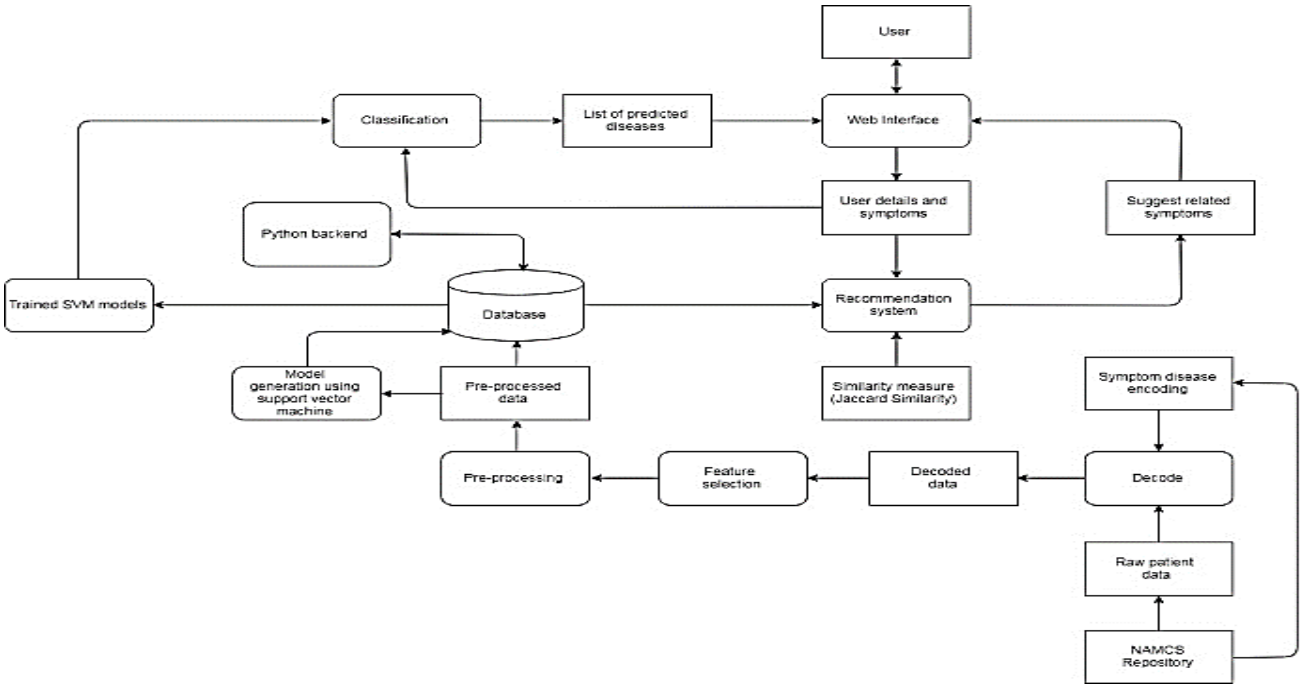

Figure 1: Proposed System Architecture

classification algorithm which works well with both numeric and categorical attributes. The goal of SVM is to find an optimal separating hyperplane which maximizes the margin for training dataset. The equation of this hyperplane is used as a hypothesis to predict the class labels for test data. The Support Vector Machine uses the linear classifier of the following form,

$$\boldsymbol{f(x) = (W \times I) + bias}$$

where W=weight factor, I=input vector and bias. The hyperplane which divides is defined by f(x)=0. Therefore the first class that falls above the hyperplane has f(x)>0 and another class below the plane is f(x)<0. The margin does not have any point in the interior region. [2] SVM model gets trained quickly as compared to artificial neural networks and has pretty good accuracy for small and medium-sized datasets. SVM algorithm is supported with different kernels for finding similarity between two feature vectors**.**

3.2 Symptom Recommendation

Simple algorithms are used by recommendation systems with an aim to provide the most accurate and relevant items to the user with help of filtering information that is useful from a large pool of information base. They discover patterns in the data present in the dataset by learning the choices of users and show results that co-relates to their needs. The proposed system recommends symptoms to the user using collaborative filtering approach in which user data is considered while processing information for the recommendation. The patient dataset is divided into separate classes based on age (less than 15, 15 to 24, 25 to 44, 45 to 64, 65 and 74 and 75 above), body mass index (bmi, less than 18.5 as underweight, 18.5 to 25 as normal, 25 to 30 as overweight, 30 to 40 as obese and more than 40 as morbidly obese), gender (male or female), tobacco use (current smoker, former smoker or never) and injury value (no, yes and trauma, yes and medical/surgical treatment or yes and overdose/poisoning). Therefore the number of classes = 6 (for age)×5(for BMI) ×2(for gender) ×3(for tobacco use) ×4(for injury) = 720 classes. These user details are collected and the user is mapped to one of 720 classes as user class. Collaborative filtering is then used to filter the patient records which belong only to this class. The user enters a symptom and patient records of the user class having this symptom are used to find other symptoms indicated by them. These related symptoms are suggested to the user. The user then might select one or more suggested symptoms based on which more related symptoms are suggested. This helps the user to look for symptoms suffered by patients with similar conditions.

3.3 Similarity measure

In order to measure the similarity between subsets of features, we use Jaccard coefficient.[4] Jaccard coefficient is a metric used for matching the similarity and diversity of two sets. It is defined as the ratio of the size of the intersection to the size of the union of the given sets. Given two sets Sq and Sq', the Jaccard coefficient is defined as follows [4]:

$$\boldsymbol{J(Sq, Sq') = |Sq \cap Sq'| \div |Sq \cup Sq'|}$$

$$\boldsymbol{such\ that\ 0 \leq J(Sq, Sq') \leq 1}$$

The model takes into consideration a total of 425 symptoms that are relevant to humans. Each patient record has five descriptive parameters namely age, BMI, gender, injury, tobacco use, which are important in determining the class of the user and 425 symptoms. Here, Jaccard coefficient is used to compare two features The patient records are then sorted in descending order of their similarity measure with user tuple. A predetermined threshold (a fraction between 0 and 1) is selected to filter out the patient records below it.Then for each patient record in filtered patients records if the patient has a symptom and if it is not selected by the user, that symptom is suggested to the user. Thus more likely symptoms are at the top with less likely at the bottom.

3.4 Python Backend

The Python backend communicates with all the other modules and acts as the server for the application. Django is an open source Python web framework that allows users to write backend scripts and it follows Model View Template(MVT) pattern. This module communicates with the disease prediction module to make predictions, and also sends and receives data from the Web interface. The major role of this module is to render the web-based user interface for the clients. User authentication, input validation, views, and forms are handled by this module

3.5 Web Interface

In order to interact with the user, the proposed system uses the web interface.which acts as a layer of abstraction between the Python backend and the user. The interface takes user details such as body mass index, gender, tobacco use and injury as input and allows the user to select one or more symptoms which he/she might be suffering.A list of related symptoms is then suggested to the user from which the user can add/remove symptoms and the suggested symptoms list get modified accordingly. Once the user is done selecting all the symptoms, a list of all probable diseases which the user might be suffering from is shown.

4. PREDICTING DISEASE

Once all symptoms are taken from the user, the user information which includes age, BMI, gender, injury, tobacco use along with the collected symptoms is given to each support vector machine disease model individually for predicting the status i.e 'Yes' if the user shows signs of that disease and 'No' otherwise.

The linear kernel is used for support vector machine in the proposed system with parameters as squared hinge loss function, penalty term as 1.0 and maximum iterations as 1000. In order to take into account, the penalty for the inaccuracy of predictions in classification problems, a computationally feasible loss function such as hinge loss function is used.Hinge loss is used as a regularization parameter.Since the goal of SVM is to maximize the margin between support vectors, the accuracy of the classifier reduces in the process since data may not be linearly separable.Hence there is always a tradeoff between maximization of margin and minimization of classification error.

5. METHODOLOGY AND EVALUATION

The proposed machine learning technique was tested on data set obtained from NHAMCS(National Hospital Ambulatory Medical Care Survey) which was available for free.

5.1 Data Description

The dataset is obtained from the NHAMCS repository which was present in form of Electronic Medical Record (EMR). It consists of over 37,000 records and a total of 430 attributes. Every tuple has 22 disease status attributes associated with it, which are used as the output labels.

5.2 Data Preparation and Preprocessing

In order to improve the quality of the predictions and to enhance the accuracy of the model, the data has to be preprocessed first by dealing with noisy and inconsistent data. In EMR data, patients' diseases and hospital interventions are captured through a set of diagnoses and procedures codes. These codes can be used as features to build a prediction model and an appropriate feature selection can inform a clinician about important risk factors for a disease[4]. The NAMCS dataset was decoded from this EMR to give a table of patient records with each record having a symptom/diagnosis and a list of diseases. A list of symptoms along with their codes from NAMCS documentation is recorded which are then used to create dummy variables with boolean values True or False. Thus, a single column for a list of symptoms for each patient is expanded to produce a separate column for each symptom with boolean values, True if the patient has that symptom or disease or False otherwise. Dummy variables are necessary so that the data can be used by a classification algorithm since multi-valued attributes cannot be used. Tuples with missing values are removed. Categorical and numerical features are pre-processed separately.

5.3 Training

Support Vector Machine(SVM) model was created and trained for each individual disease from a selected list of commonly occurring diseases. The kernel used here is linear kernel since the number of features is large and the linear kernel is less prone to overfitting as opposed to polynomial, radial basis or Gaussian kernel. An 80-20 split was performed to get the training data and the testing data, where 80% of the data was used for training and 20% was used for testing. The data was shuffled randomly before performing the split.

6. RESULTS

The bar plot in Figure. 2 below shows the comparison of accuracy on test data for three diseases viz. Arthritis, Hypertension, and Cancer. It can be seen that the accuracy of Support Vector Machine is better than decision tree and Naïve Bayes and is comparable to that of random forest.

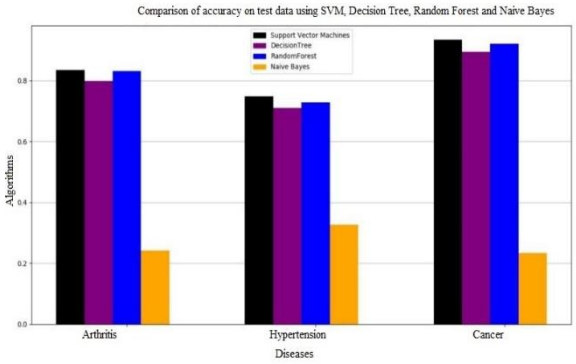

Figure 2: Comparison of accuracy on test data for SVM, Decision Tree, Random Forest, Naïve Bayes classifiers

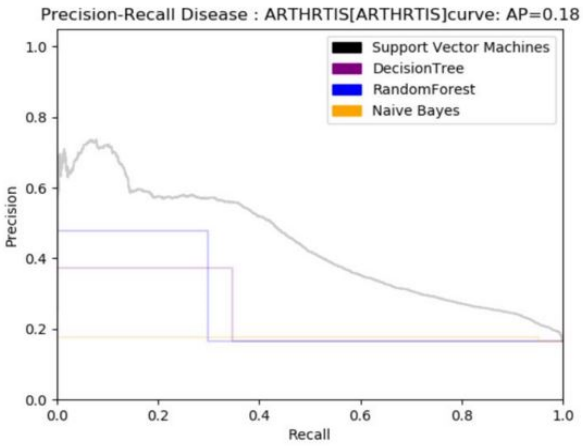

Figure 3: Precision-Recall curve for disease 'Arthritis'

The above figure (Figure. 3) shows the precision-recall curve for disease 'Arthritis' using algorithms Support Vector Machine, Decision Tree, Random Forest, Naive Bayes.The ideal precision-recall (PR) curve is a straight line with equation precision=1. Here we can observe that PR curve for SVM is closer to the ideal curve than decision tree, random forest or Naive Bayes.

Table 1,2,3 shows the training results obtained for 3 diseases viz. Arthritis, Cancer, Hypertension respectively.

Table 1. Training results for the disease Arthritis on the proposed model

Algorithm	Training accuracy	Testing accuracy
SVM	85.73 %	84.85 %
Decision tree	97.67%	79.93 %
Random forest	96.39 %	83 %
Naïve Bayes	26.4 %	25.41 %

Table 2. Training results for the disease Cancer on the proposed model

Algorithm	Training accuracy	Testing accuracy
SVM	93.74 %	93.35 %
Decision tree	89.42%	89.42 %
Random forest	97.45 %	92.22 %
Naïve Bayes	24.55 %	22.47 %

Table 3. Training results for Hypertension

Algorithm	Training accuracy	Testing accuracy
SVM	79 %	75.3 %
Decision tree	96%	71.65 %
Random forest	94.58 %	73.79 %
Naïve Bayes	33.8 %	32.5 %

From the above tables, we can see that support vector machine show almost same accuracy for training and test set whereas decision tree and random forest show high accuracy on training set which differs from a large margin with respect to test set thus indicating high variance or overfitting. Naive Bayes shows poor performance on both training and test data.

The tables 4,5,6,7 given below show the confusion matrices for the algorithms Support vector machine, Decision Tree, Random Forest and Naive Bayes on disease 'Arthritis' along with precision and recall values respectively

Table 4: Confusion matrix for Support Vector Machine

Actual-> Observed	**No**	**Yes**
No	6057	986
Yes	168	276

Precision: 0.86

Recall: 0.97

Table 5: Confusion matrix for Decision Tree

Actual-> Observed	**No**	**Yes**
No	5639	586
Yes	842	420

Precision: 0.87

Recall: 0.91

Table 6: Confusion matrix for Random Forest

Actual-> Observed	**No**	**Yes**
No	5827	398
Yes	868	394

Precision: 0.87

Recall: 0.93

Table 7: Confusion matrix for Naive Bayes

Actual-> Observed	**No**	**Yes**
No	760	5465
Yes	58	1204

Precision: 0.92

Recall: 0.12

The above tables indicate that SVM has marginally high precision and recall score as compared to other algorithms. Naive Bayes has a good precision score but poor recall.

7. CONCLUSION

An efficient technique for disease prediction using support vector machines was presented in this paper. The paper also provided an explanation of how collaborative filtering can be used to create a recommendation system for suggesting related symptoms.The system used Jaccard similarity measure for collaborative filtering. Various classification algorithms were compared using Precision-recall curve and accuracy on training and test set was calculated and support vector machines were then chosen. The statistical details for the same have also been presented.

The proposed system performs disease prediction based on recent symptoms the person is suffering but doesn't take into consideration the duration for which the person is suffering these symptoms. Effective results can be obtained if patients health profile generated overs years is used for comparison and giving weights according to these durations.

8. FUTURE WORK

The proposed system gives a fairly accurate prediction for a list of commonly occurring but serious diseases. For future work, one can rather focus on creating a single prediction model instead of a model for every disease. With the advent of health monitoring devices, one can produce a system providing an automatic and instant status of health report for an individual user without requiring user intervention or input to the system and also taking into consideration living conditions, habits and user history.Another feature which can be added to the system is the recommendation of doctors. Based on the symptoms entered by the patients the system will suggest specialists who are close to the patient's location area.

9. ACKNOWLEDGEMENTS

We would like to thank our mentor Mrs. Rajni Pamnani, who is an assistant professor at KJ Somaiya College of Engineering, for lending her support and expertise, which were instrumental in giving our project a basis and direction.

10. REFERENCES

[1] Portugal, Ivens, Paulo Alencar, and Donald Cowan. "The use of machine learning algorithms in recommender systems: a systematic review." Expert Systems with Applications (2017).

[2] Radhimeenakshi, S. "Classification and prediction of heart disease risk using data mining techniques of Support Vector Machine and Artificial Neural Network." Computing for Sustainable Global Development (INDIACom), 2016 3rd International Conference on. IEEE, 2016.

[3] Kulev, Igor, et al. "Recommendation algorithm based on collaborative filtering and its application in health care." (2013): 34-38.

[4] Kamkar, Iman, et al. "Stable feature selection for clinical prediction: Exploiting ICD tree structure using Tree-Lasso." Journal of biomedical informatics 53 (2015): 277-290.

[5] Mahajan, Shashank, and Gaurav Shrivastava. "Effective Diagnosis of Diseases through Symptoms Using Artificial Intelligence and Neural Network." International Journal of Engineering Research and Applications: 2248-962.

[6] Prokosch, Hans-Ulrich, and Thomas Ganslandt. "Perspectives for medical informatics." Methods of information in medicine 48.01 (2009): 38-44.