Abstract— Medical data is a source of information generated from the hospitals in the record form. The system proposed here uses a vast storage of information. The diagnosis is made based on this historical data. It focuses on computing the probability of occurrence of a particular ailment from the medical data by mining it using a unique algorithm. It combines the key points of Large Memory Storage, and Retrieval, Naive Bayes, and differential diagnosis all integrated into one single algorithm. This algorithm can be used in solving a few common problems that such as diagnosis of multiple diseases showing similar symptoms, to get faster second opinion, and to get accurate trends present in the medical records.

Keywords— Medical diagnosis, data mining, differential diagnosis, neural networks, knowledge discovery.

I. INTRODUCTION

Since the advent of advanced computing, doctors have used technologies such as surgical imagery to X-ray photography. Diagnosis is a process that still requires doctor’s knowledge and experience. Some factors that affect diagnosis are medical history, climatic conditions, blood pressure, and so on. Medical decision support systems are essential and help doctors in taking correct decisions. Medical decision is challenging job in case of diseases with similar symptoms, or in case of diseases that occur rarely. Misdiagnosis is a problem and can occur due to inexperiance of the doctors, repetitive diagnosis by experienced doctors, incomplete information, stress, fatigue, and so on. In developing countries, lack of trained and experienced doctors is a major problem. In the existing methodology two databases are used that is Disease/Symptoms database and the Records Database. And the Disease/Symptom database contains the list of the existing known diseases and their corresponding symptoms along with weights. Consider the case of a description database. The database obtained is a description database which contains diseases and their corresponding symptoms in the form of a description. The representation and quality of data is first. So to increase the accuracy the relevant information must be retrieved from this database. Then retrieved features are weighted. Searching a database of diseases and symptoms is time consuming thereby requires large database access which decreases the accuracy of the system.

The relevant informations are retrieved from description database using tokenization, filtering and stemming. Tokenization is the process of breaking up a text into words, phrases, symbols called tokens and at the same time throwing away certain characters, such as punctuation. Filtering means content filtering that is removing the nouns, prepositions, etc from the database. Stemming is the process for reducing inflected words to their stem, base or root form. This described method provides a more efficient way to obtain high quality data.

The naive bayes classifier is a good tool in medical diagnosis. For example given a list of symptoms, it predicts occurrence of a disease. The classifier computes the probability of each attribute. NB's main strength is its simplicity, efficiency and good performance. With this new database the root disease and the next probable diseases can be obtained. Symptom matching, naive bayes, differential diagnosis and LAMSTAR networks are integrated into one algorithm. Symptom matching and differential diagnosis are used to obtain the root disease and next probable disease. With the naive bayes the weights of symptoms are calculated. A method is proposed to decrease the database access. The diseases corresponding to symptom with least weight is considered and checks whether other entered symptoms are present in that diseases thereby reducing the database access. The system will in turn display the root disease and next probable diseases with similar symptoms. The list of diseases that has maximum probability of suffering from is obtained. With this the doctors can recommend specific tests to undergo thus reducing the number of tests thereby saving time and money.

II. RELATED WORKS

There are many ongoing researches in the field of medical diagnosis. Iliad [1] is an expert system with Bayesian reasoning to calculate the posterior probabilities of various diagnoses. DXplain [2] is a decision support system used to produce a ranked list of diagnoses. Bayesian networks (BN) plays an important role in medical diagnostics. It is used to represent conditional dependencies among the random variables. It computes the probability of the occurrence of various diseases when the symptoms are given [3]. It contains nodes and edges forming directed acyclic graph. With this nodes and edges it is easily understandable when compared to other techniques. BN have both quantitative and qualitative part. The statistics to fill the numeric information are not available. The KNN is the simplest of all classifiers and is used in predicting diseases. For classification majority vote is considered. Object is assigned to class with highest match. As
number of classes increases the performance of KNN increases [4]. The number of neighbours are obtained with value of k [5]. Implementation of KNN mechanism is easy and the debugging process is very faster. As value of k decreases noise points in training set increases. As value of k increases it becomes expensive [6]. ANN also called Neural Network consists of neurons that are interconnected. They represent relationships between inputs and outputs. Patient is assigned to one of the classes of diseases with this network. So ANN is a powerful tool for disease diagnosis [7]. ANN is good in identifying diseases and does not need any details of how to recognize as it learns by example [8]. It is easy to maintain and has good capacity. Has good computational power with good accuracy. It is like a black box that is there is no knowledge of internal workings with which network solution can not be interpreted [9].

Back propagation is used to train artificial neural networks. It is a supervised learning method. The connection between the layers has weights and is trained using back propagation learning algorithm [10]. They are widely used in diagnosis of diseases. Only small amount of training data is required and works better in real world applications [11]. The converging of result to local minimum is very slow in this network and it is not guaranteed too. Assigning of weight is a slow technique. Support Vector Machine (SVM) is a supervised learning method. It takes input and predicts to which of the class it belongs. The input is considered as points in this feature space. These points are separated using the hyperplane [12]. Its analysis is simple and easy [13]. Performance is good and it delivers a correct solution. It can be used in classification and prediction. Size and speed is another limitation of SVM. The transparency of results is a shortage in SVM.

III. PROPOSED METHODOLOGY

The system has a three-tier workflow pattern. The method for triple precision diagnosis is shown in Fig. 1.
The first step is to reduce the description dataset to improve the quality of data. So for that tokenization is done after that filtering is done to remove the nouns, propositions and so on. Then stemming is done using Porters Stemmer algorithm. Thus obtain the reduced dataset. Then it involves entering the symptoms given by the patient. As an output, the system gives the list of all possible diseases ranked according to the number of symptoms matched in the database. The list is generated after entering the symptoms. The ranking is generated according to the percentage match of the total number of symptoms entered. If a single disease in the given subset gains maximum weight above all other diseases, it is the interpreted by the system as the possible diagnosis. Instead, if multiple diseases are found with nearest weights or same weights, then the system proceeds to mining medical records.

The second step involves mining medical records. If multiple diseases are found with similar ranking, it becomes difficult to pinpoint to one of them. This is especially the case of some diseases in an area, or some form of rare disease, or disease occurring due to various other factors. In such cases, it becomes very difficult to point at one disease using the symptom matching method. In such cases, recent medical historical data stored in the database of the proposed system is used, within an interval of three months to list the diseases occurred. If the obtained diagnosis is still vague, then the system goes to third step that is the differential diagnosis. And thus generates the list of diseases occurred more accurately.

IV. IMPLEMENTATION

Various algorithms utilized in this method are explained below. The system uses a three-tier workflow method for triple precision diagnosis. Tokenization means breaking a text into words, symbols, or elements called tokens. This token is the input for further processing such as parsing. The alphabetic characters are part of a token and are separated by whitespace characters such as space, punctuation or line break. Filtering means content filtering that is removing the nouns, prepositions, etc from the database. The Porter stemming algorithm is a process to bring a word to its stem form. Removing suffixes by automatic means is an operation which is especially useful in medical diagnosis. With which the performance can be improved. The suffix stripping process will reduce the terms in the system thereby reducing the size and complexity of the data in the system.

![Porter Stemmer Flowchart](image)

The Porter Stemmer is a linear step stemmer. It has five steps and rules are applied to each step. In each step, the conditions attached to a rule are tested when a word is matched to that suffix rule, to know what would be the resulting stem, when the corresponding suffix is removed. Once the condition is satisfied the rule fires and the corresponding suffix will be removed and then control moves to following step. The next rule in the step is tested when the first rule is not accepted and it continues this process until there are no more rules in that step or a rule from that step fires and control passes to the next step. This process continues for every step, and the stemmer returns the resultant stem after control is passed from fifth step.
After retrieving the relevant information the reduced dataset is obtained. Now it involves entering the symptoms given by the patient. The system provides the list of all possible diseases according to the number of symptoms matched in the database. After entering the symptoms the weights are calculated. Weight is calculated using the naïve bayes concept. Weight is the quotient of number of occurrence of symptom to total number in the dataset. A method is proposed to decrease the database access. The diseases corresponding to symptom with least weight is considered and checks whether other entered symptoms are present in that diseases thereby reducing the database access. If a single disease is obtained with maximum weight then it is the possible diagnosis. If multiple diseases are obtained with same or nearest weights, then the system proceeds to mining medical records. For example, flu epidemic has shown the same symptoms as that of viral fever initially, resulting in multiple diseases with similar symptoms. So it becomes very difficult to point at one disease. In such cases, historical data stored in the database is used, with an interval of three months. The system lists the diseases occurring in the review period. The three months time frame gave accurate diagnosis. If the obtained diagnosis is still vague, then the system uses differential diagnosis. For this system uses Hopfield networks. It can be used to find the relative frequencies. A differential diagnosis is a diagnostic method to identify an entity whose multiple alternatives are possible. It lowers probabilities of candidate conditions to negligible levels.

When the new input word is presented to the system the LAMSTAR network inspects all weights stored. The correct symptom is obtained from dataset and then it is compared with the original symptoms entered. If any stored pattern matches the input word, the system updates weights according to the following procedure:

\[ W(t+1) = W(t) + a[X(t) - W(t)] \]

Where \( W(t+1) \) is the modified weight, \( W(t) \) is the existing weight and \( a \) is the learning coefficient. If no match was found, the system stores new pattern in the dataset of diseases and symptoms. It is faster in modifying the weights. It is very useful in the field of medical diagnosis.

V. EXPERIMENTAL RESULTS AND ANALYSIS

A prototype system using web tools and technologies like PHP and MySQL was developed. The data have been obtained from limited scope. The system was run for a dataset of 4000 diseases. The dataset contains diseases along with their symptoms. Specific test cases were run, and the following results were obtained. The dataset used here is a description dataset. The quality and representation of data is important. So to reduce the dataset tokenization, filtering, and stemming is done to the dataset. It is done to the entire dataset. For example in case of disease "Acute Bronchitis" the symptoms field contain the description "Symptoms includes coughing with sputum, shortness of breathing or dyspnea, wheezing, chest pains, fatigue and malaise" which after filtering and stemming becomes "include cough sputum short of breath dyspnea wheezes chest pain fatigue malaise". This increases the quality of data and reduces the database size. Fig. 3 depicts the number of symptoms matching and the graph shows the probability of occurrence of diseases. The graph shows the probability of occurrence of diseases.
The system shows the list of diseases found by matching symptoms and its probabilities of occurrence is calculated. The second step resulted in very accurate prediction of diseases based on the recent trends. It accurately caught influenza and malaria for the given symptoms, during the review period of three month. This diagnosis matched very accurately with the patient’s actual ailment.

Fig. 4 shows the same result, but with more priority given to diseases occurring within a three month period. The graph shows the probability of occurrence of diseases.

**Last 3 Month Data**

<table>
<thead>
<tr>
<th></th>
<th>Eight Day Measles</th>
<th>Influenza</th>
<th>Sore Throat</th>
<th>Whooping Cough</th>
<th>Malaria</th>
<th>Pharyngitis</th>
</tr>
</thead>
<tbody>
<tr>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>102</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Fig. 5 shows the list of diseases and their probabilities of occurring calculated on the basis of differential diagnosis technique. The graph shows a graphical representation of the same. In this system, an additional step, where differential diagnosis has been combined with recent medical history, to get more accurate results, has been implemented. This step has given accurate results to catch recent trends.
A. Performance Analysis

Initially the system using KNN, differential diagnosis, and LAMSTAR techniques was implemented. But by using naive bayes, the overall speed and accuracy of the algorithm increased considerably. Especially in case of larger datasets, LAMS TAR gave faster and better results. To undergo performance analysis a sample dataset of diseases along with symptoms are considered. Speed tests were performed by calculating the total execution time taken for the weight calculation. For this the microtime() function of PHP is used in case of KNN and Naive bayes. The improvement in speed was contributed due to faster calculation of weights by Naive bayes.

The implementation of KNN is easy but there is some difficulty in using KNN. Increased value for k reduces the noisy points in the training data set. As k increases it becomes computationally expensive and can also increase the computational complexity and time. As all process is done at run time it is comparatively a slow technique. The better enhancement to this is the use of naive bayes classifier. Naive bayes algorithm outperforms KNN in weight calculation. It is a good tool in medical diagnosis. For example given a list of symptoms, it predicts occurrence of a disease. The classifier computes the probability of each attribute. NB's main strength is its simplicity, efficiency and good performance. It combines efficiency with good accuracy. Due to its good accuracy it is used in medical diagnosis. A small set of training data is required for the estimation of variable values. NB is a very powerful technique in diagnosing diseases. It is used to provide efficient output. The NB classifier needs a very large training set to obtain good results.

<table>
<thead>
<tr>
<th>TABLE I</th>
<th>SPEED TEST DONE IN WEIGHT CALCULATION</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Test(Sec)</td>
</tr>
<tr>
<td>K Nearest Neighbour</td>
<td>35.77</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>9.958</td>
</tr>
</tbody>
</table>

VI. Conclusion

Medical diagnosis is an important area of research which helps to identify the occurrence of a disease. Medical data is an ever growing source of information. The system, making use of various techniques mentioned, will in turn display the root disease along with the set of most probable diseases which have similar symptoms. The database used is a
description database so to reduce the dataset tokenization, filtering and stemming is done. In this system, by using differential diagnosis, LAMSTAR, and Naïve bayes, an attempt has been made to assist the doctors to perform diagnosis in accurate way. The main advantage of the system is that it can be applied to any kind of dataset whether it is a description dataset or not.

Uncertainty is a factor in medical diagnosis as the system is built on experience of doctors. The systems results are promising. It cannot be used as a substitute to diagnosis, it helps doctors to reach in a solution. The system gives list of diseases with which the doctors can specify the corresponding test thereby saving the time and money. A comparison between naïve bayes and KNN is done in weight calculation. Naïve bayes outperforms KNN in weight calculation. In the future along with finding the probable diseases the medications and test required can be incorporated which increases the accuracy of the system. More intelligence can be incorporated to the system with which the accuracy can be increased.

ACKNOWLEDGMENT

The authors would like to thank Mr. Arun Kumar for the expertise and medical domain guidance and would also like to convey special thanks for his constant support.

REFERENCES


[5] Rahul Isola, Student Member, IEEE, Rebeck Carvalho, Student Member, IEEE, and Amiya Kumar Tripathy, Member, IEEE, Knowledge Discovery in Medical Systems Using Differential Diagnosis, LAMSTAR, and KNN, IEEE Transactions on Information Technology in Biomedicine, Vol. 16, No. 6, November 2012.


