A survey on utilization of data mining approaches for dermatological (skin) diseases prediction

Barati, E, Saraee, MH, Mohammadi, A, Adibi, N and Ahmadzadeh, MR

Title	A survey on utilization of data mining approaches for dermatological (skin) diseases prediction
Authors	Barati, E, Saraee, MH, Mohammadi, A, Adibi, N and Ahmadzadeh, MR
Type	Article
URL	This version is available at: http://usir.salford.ac.uk/id/eprint/18594/
Published Date	2011

USIR is a digital collection of the research output of the University of Salford. Where copyright permits, full text material held in the repository is made freely available online and can be read, downloaded and copied for non-commercial private study or research purposes. Please check the manuscript for any further copyright restrictions.

For more information, including our policy and submission procedure, please contact the Repository Team at: usir@salford.ac.uk.

A Survey on Utilization of Data Mining Approaches for Dermatological (Skin) Diseases Prediction

E. Barati, M. Saraee, A. Mohammadi, N. Adibi and M. R. Ahamadzadeh

***Abstract*— Due to recent technology advances, large volumes of medical data is obtained. These data contain valuable information. Therefore data mining techniques can be used to extract useful patterns. This paper is intended to introduce data mining and its various techniques and a survey of the available literature on medical data mining. We emphasize mainly on the application of data mining on skin diseases. A categorization has been provided based on the different data mining techniques. The utility of the various data mining methodologies is highlighted. Generally association mining is suitable for extracting rules. It has been used especially in cancer diagnosis. Classification is a robust method in medical mining. In this paper, we have summarized the different uses of classification in dermatology. It is one of the most important methods for diagnosis of erythemato-squamous diseases. There are different methods like Neural Networks, Genetic Algorithms and fuzzy classifiaction in this topic. Clustering is a useful method in medical images mining. The purpose of clustering techniques is to find a structure for the given data by finding similarities between data according to data characteristics. Clustering has some applications in dermatology. Besides introducing different mining methods, we have investigated some challenges which exist in mining skin data.**

Index Terms— Erythemato-squamous diseases, Data mining, Dermatology, Medical data mining.

I. INTRODUCTION

THE interest in systems for autonomous decisions in medical and engineering applications is growing, as data becomes easily available.

In the last century, an exponential inhancement has been seen in the accuracy and sensitivity of diagnostic tests, from observing external symptoms and using sophisticated laboratory tests and complex imaging methods increasingly that permit detailed non-invasive internal examinations. This improved accuracy has inevitably resulted in an exponential increase in the patient data available to the physician. The process of finding evidence to distinguish a probable cause of patient's key symptoms from all other possible cause of the symptom are known as establishing a medical diagnosis.

The use of computer technology in medical decision support is now widespread and pervasive across a wide range of medical area, such as cancer research, dermatology, etc. Data mining is a tremendous opportunity to assist physician deal with this large amount of data. Its methods can help physicians in various ways such as interpreting complex diagnostic tests, combining information from multiple sources (sample movies, images, clinical data, proteomics and scientific knowledge), providing support for differential diagnosis and providing patient-specific prognosis.

The rest of the paper is organized as follows: It first gives an introduction to data mining and its different methods. Then medical data mining is described. In section IV, some instances of the prediction and diagnosis problems in medicine in case of skin diseases are considered. The article ends by concluding with a summary of investigated methods and future research.

II. WHAT IS DATA MINING?

Simply stated, data mining refers to *extracting or "mining" knowledge from large amounts of data or databases* [1]. The process of finding useful patterns or meaning in raw data has been called KDD (knowledge discovery in databases) [2]. Knowledge discovery consists of an iterative sequence of the following steps: Data cleaning to remove noise and inconsistent data, Data integration to combine multiple data sources, Data selection for retrieving data relevant to the analysis task from the database, Data transformation for transforming data into forms appropriate for mining by performing summery of aggregation operation, Data mining which is an essential process of applying intelligent methods in order to extract data patterns, Pattern evaluation to identify the truly interesting patterns based on some interestingness measures and knowledge presentation which uses visualization and knowledge representation for presenting the mined knowledge to the user [1].

March 10, 2011,

E. Barati is with Intelligent Databases, Data mining and Bioinformatics Laboratory, Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran. (email: e.barati@ec.iut.ac.ir)

M. Saraee is with Intelligent Databases, Data mining and Bioinformatics Laboratory, Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran. (email: sareee@cc.iut.ac.ir)

A. Mohammadi is with Intelligent Databases, Data mining and Bioinformatics Laboratory, Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran. (email: amohammadi@ec.iut.ac.ir)

N. Adibi is a researcher of skin disease and leishmaniasis research center of Isfahan University of Medical Science, Isfahan, Iran.(email: nedaadibi705@gmail.com)

M. R. Ahmadzadeh is with Department of Electrical and Computer Engineering, Isfahan University of Technology, Isfahan, Iran. (email: ahmadzadeh@cc.iut.ac.ir)

There are various numbers of data mining methods. One approach to categorize different data mining methods is based on their function ability as below [3]:

1) ***Regression*** is a statistical methodology that is often used for numeric prediction.
2) ***Association*** returns affinities of a set of records.
3) ***Sequential pattern*** function searches for frequent *subsequences* in a *sequence dataset*, where a sequence records an ordering of events.
4) ***Summarization*** is to make compact description for a subset of data.
5) ***Classification*** maps a data item into one of the predefined classes.
6) ***Clustering*** identifies a finite set of categories to describe the data.
7) ***Dependency modeling*** describes significant dependencies between variables.
8) ***Change and deviation detection*** is to discover the most significant changes in the data by using previously measured values.

III. MEDICAL DATA MINING

We live in data-rich times and each day, more data are collected and stored in databases. Increasing the use of data toward answering and understating important questions has driven the development of data mining techniques. The purpose of these techniques is to find information within the large collection of data. Although data mining is a new field of study of medical informatics, the application of analytical techniques to discover patterns has a rich history. Perhaps it was one of the most successful uses of data analysis for discovering and understanding of the medical science, especially infectious disease [4].

Medical diagnosis is known to be subjective and depends not only on the available data but also on the experience of the physician and even on the psycho-physiological condition of the physician. A number of studies have shown that the diagnosis of one patient can differ significantly if the patient is examined by different physicians or even by the same physician at various times [5].

The idea of medical data mining is to extract hidden knowledge in medical field using data mining techniques. One of the positive aspects is to discover the important patterns. It is possible to identify patterns even if we do not have fully understood the casual mechanisms behind those patterns. In this case, data mining prepares the ability of research and discovery that may not have been evident. Even the patterns which are irrelevant can be discovered [4]. Clinical repositories containing large amounts of biological, clinical, and administrative data are increasingly becoming available as health care systems integrate patient information for research and utilization objectives [6]. Data mining techniques applied on these databases discover relationships and patterns which are helpful in studying the progression and the management of disease [7]. A typical clinic data mining research including following ring: structured data narrative text, hypotheses, tabulate data statistics, analysis interpretation, new knowledge more questions, outcomes observations and structured data narrative text [8]. Prediction or early diagnosis of a disease can be kinds of evaluation. About diseases like skin cancer, breast cancer or lung cancer early detection is vital because it can help in saving a patient's life [9].

In order to predict unknown or future values of interest, prediction can be used, whereas description focuses on finding patterns describing the data and the subsequent presentation. For instance, one may classify diseases and provide the symptoms, which describe each class or subclass [5].

Healthcare related data mining is one of the most rewarding and challenging areas of application in data mining and knowledge discovery. The challenges are due to the datasets which are large, complex, heterogeneous, hierarchical, time series and varying of quality. As the available healthcare datasets are fragmented and distributed in nature, thereby making the process of data integration is a highly challenging task [10].

IV. THE APPLICATION OF DATA MINING IN HEALTHCARE: IN CASE OF SKIN DISEASES

Recently, skin diseases have been common to everyone. Many factors influence the onsets of these diseases and each age group usually has its different symptoms. Growing bacteria and molds in humid, damp and hot weather conditions, also exposures to excess amounts of ultraviolet radiations in the sunlight will make skin sensitive, easy to be infected, and possibly cause skin problems. In addition to the external infections, internal sebaceous glands, dead skin, sweats, mixed with dusts and other unwanted secretions can cause other serious skin diseases. Although skin diseases are detected easier than other diseases and diagnosing symptoms and deciding treatment plans are not as complex as other internal diseases, many people often ignore the importance of them.

With the rapid advancement in information technology, many different data mining techniques and approaches have been applied to complementary medicine. Statistics provide an impressive background to define and evaluate the result.

Here, we intend to express some of data mining applications deal with skin disease. Before that preprocessing and variable selection is described.

A. Preprocessing and variable selection

Reduction of feature dimensionality is of considerable importance in data mining. The reason for being so is twofold: to reduce the computational complexity and to improve the classifier's generalization ability. Feature extraction and feature selection are two different approaches for the reduction of dimensionality. Feature extraction involves linear or nonlinear transformation from the original feature space to a new one of lower dimensionality. Feature selection, on the other hand, directly reduces the number of original features by selecting a subset of them that still retains sufficient information for classification. Feature selection plays an important role in dimension reduction for classification [11]. It can not only reduce the dimension of data, but also lower the computation consumption, and so that it can gain classification

performance. The feature selection algorithms designed with different evaluation criteria are divided into two categories: the filter methods and wrappers methods [12] [13]. The filter methods mainly identify a feature subset from the original feature set with a given evaluation criterions which are independent of learning algorithm. The wrapper methods choose those features with high prediction performance estimated by specified learning algorithms.

B. Association

Data mining is useful for generating hypothesis for further testing as in identifying associations or relationships between data that are then tested using conventional statistical techniques [14]. Association rule mining is one of the major data mining techniques used to discover novel and interesting relationships among data objects present in databases [1]. Association mining consists of two important steps, namely frequent patterns discovery and rule construction [15]. Frequent pattern discovery is considered as the major step in the process and the one that presents the greatest difficulty in terms of space and time complexity. A frequent itemset usually corresponds to a set of items that occur frequently together within a transactional database.

Association rule mining has been widely used for discovering rules in medical applications. Three challenges of association rule mining approaches in these applications are 1) most widely used interestingness criteria, such as confidence and lift, do not make sense to medical practitioners, 2) too many trivial rules discovered overwhelm truly interesting rules and 3) an association rule mining approach is inefficient when the frequency requirement. The minimum support is set low [16].

Wang and Chung [17], made a web-based data browsing and content-based retrieval system for a skin cancer database. A skin cancer image database was created using a three-tier system. For this skin cancer database, one table was designed to store features of the skin tumors. Besides the tumor features, some other attributes were added into the table. These included record number, patient id number, the date that the image was taken, the image id and the image file name. This database has been designed to be accessible through the Internet. It makes it easy for people to browse, query, visualize images, and get data mining results. Various browsing and content-based retrieval methods are supported for the skin cancer image database through web-based graphic user interfaces. Since each attribute has a value in every record in the skin cancer database, association rules are based on which attribute value ranges frequently appear together in the database. There exist a few methods to partition each attribute value range. In this study [17], Equal-population was used. It means the total value range is divided into a certain number of intervals, such that almost same numbers of records are in each interval regardless of the size of the interval.

The SETM algorithm [18] was also used to find association rules between different features of the skin cancer images. The SETM algorithm was motivated by the desire to use SQL to compute large itemsets. It generates candidates on the fly based transactions read from the database. However, to use the standard SQL join operation for candidate generation, SETM separates candidate generation from counting [18].

In the skin cancer database, the number of intervals is decided by the user while the minimum and the maximum number of intervals are predefined. In the next step frequent itemsets are found and the next stage is to use the frequent itemsets to generate the desire association rules. Finding association rules among attributes is totally dependent on the end user. The interface was designed such that user can find association rules among any combination of the attributes. Users can find association rules between different skin cancer feature values, which can be very useful for skin cancer diagnosis and study.

In [19], the significant risk factors were identified for particular types of cancer. In this study, a risk factor dataset containing patient records with categorical attributes was first constructed. Then three association rule mining algorithms, Apriori [15], Predictive apriori [20] and Tertius algorithm [21] were employed. The idea of using these algorithms is, discovering the most significant risk factors for particular types of cancer to show the highest confidence values. By applying association rule mining algorithms, in addition to finding all combination of the risk factors, the risk factors from frequent risk factors were used to generate the desired rules. Among the three association learning algorithms, the Apriori was the best one to extract useful rules for cancer dataset. However, in the lift measure, skin cancer showed the worst lift values. Lift quantifies the relationship between the body and the head of a rule. Basically, a lift value greater than 1 provides strong evidence that the body and the head of a rule depend on each other. A lift value below 1 state the body depends on the absence of the head or vice versa. Apriori performance was also verified with leverage measure and the most reliable performance was shown with the skin type cancer dataset. Leverage and lift measure similar things, except that leverage measures the difference between the probability of co-occurrence of the body and the head and the independent probabilities of each of the body and the head. Overall, it is suggested to use Apriori algorithm for such type of task.

C. Classification

Classification is defined as the process of discovering a model or a function that describes and distinguishes the data classes or concepts. The model discovery is achieved by analyzing a supervised dataset. The learned model then can be used to predict the classes of future data instances for which the class label is unknown, and this is often referred to as the predictive task [1].

Classification involves the need to find rules that can partition the data into different groups. In health care domain, this type of data mining technique would be important in diagnostic and treatment assistance decision making [22]. In a medical domain the supervised dataset contains past patient information with a disease based on the patient's symptoms. The learned knowledge can then be used to aid the medical expert in assessing the risk of future patients having that disease according to their symptoms [1].

The differential diagnosis of erythemato-squamous diseases is a difficult problem in dermatology. They all share the clinical features of erythema and scaling with very little differences. The diseases in this group are *psoriasis*, *seboreic dermatitis*, *lichen planus*, *pityriasis rosea*, *chronic dermatitis* and *pityriasis rubra pilaris* [23]. There have been several studies reported on erythemato-squamous diseases diagnosis. These studies have applied different methods to the given problem and achieved different classification accuracies. Among these studies the first work on the differential diagnosis of erythemato-squamous diseases was conducted by Guvenir et al. [23]. They developed a new classification algorithm, called VFI5 (for Voting Feature Intervals) and applied it to problem of differential diagnosis of Erythemato-Squamous diseases. The VFI5 classification algorithm is an improved version of the early VFI1 algorithm [24]. It represents a concept description by a set of feature intervals. The classification of a new instance is based on a voting among the classification made by the values of each feature separately. All training examples are processed at once. The VFI5 algorithm constructs intervals for each feature from the training examples. For each interval, a single value and the votes of each class in that interval are maintained. Thus, an interval may represent several classes by sorting the vote for each class. In this study, the domain contains records of patients with known diagnosis. Given a training set of such records the VFI5 classifier learns how to differentiate a new case in the domain. In this study, the dataset for the domain contained 366 instances with 34 attributes with 12 clinical features and 22 histopathological features. All of these instances were first used to obtain a description of the domain. The description consists of the feature intervals constructed for each feature. Since each feature was processed separately, the missing feature values that may appear both in the training and test instances were simply ignored in VFI5. The VFI5 algorithm achieved 96.2% accuracy on the Dermatology dataset with 22 histopathological features.

Guvenir and Emeksiz [25] presented an expert system for differential diagnosis of erythemato-squamous diseases incorporating decision made by three classification algorithms: nearest neighbor classifier [26], naïve Bayesian classifier [27] and voting feature intervals-5 [23]. The nearest neighbor classification is based on the assumption that examples which are closer in the instance space are of the same class. Nearest neighbor algorithm assumes that a new test instance belongs to the same class as its nearest neighbor among all stored training instances [26]. The used dataset in this study contains name and surname of patients, the doctor's prediction about the disease which ranges from 1 to 6 (each reflecting the label of the six eythemato-squamous diseases). Family history feature had the value 1 if any of these diseases had been observed in the family and 0 otherwise. The age feature representes the age of the patient. Every other feature (clinical and hsitopathological) was given a degree in the range of 0-3. There were 366 instances. The features of a patient were presented as a vector of features, which has 34 entries for each feature value. In this study for implementation of the nearest neighbor classification algorithm, the train data features and class values were stored in two separate arrays. All the features have linear values. The Euclidean distance metric was used to obtain the distance between two instances in the nearest neighbor classification algorithm. Also, the features were assigned weights such that the irrelevant features had lower weights while the strongly relevant features were given higher weights. Giving different weights to each feature modifies the importance of the feature in the classification process such that a relevant feature becomes more important than a less relevant one. A genetic algorithm was used to determine the weights of features to be used with the nearest neighbor classification algorithm. The same genetic algorithm was applied to determine the weights of the features to be used with the VFI5 algorithm. They found that the features acanthosis, follicular horn plug, munro microabcess and age were the least relevant.

Bayesian classifier is an algorithm that approaches the classification problem using conditional probabilities of the features [27]. The aim of Guvenir and Emeksiz [25] was to classify a single test instance depending on the previously established training dataset. The last algorithm was VFI5 [23] which constructs intervals for each feature. By using voting feature intervals -5 and 10-fold cross validation, they obtained 99.2% classification accuracy on the differential diagnosis of erythemato-squamous diseases. This expert system is a visual tool for differential diagnosis of erythemato-squamous diseases for both dermatologists and students studying dermatology. It stores the patient records in a database for future reference. Students studying dermatology can use this expert system to test their knowledge by comparing their predictions with classifications done by the algorithms. It also can be used to be a guide to the doctors in making their own diagnosis mechanisms by examining the working methodologies of the three classification algorithms.

Nanni [28] developed a new ensemble of support vector machines (SVM) [29] based on Random Subspace (RS) [30] and feature selection. SVM is primarily a dichotomy classifier. It is based on Vapnik–Chervonenkis (VC) dimension and structural risk minimization (SRM), which has been shown to be superior to any other traditional empirical risk minimization principal (ERM). SVM can handle a nonlinear classification efficiently by mapping samples from low dimensional input space into high dimensional feature space with a nonlinear kernel function. In feature space, SVM tries to maximize the generalization problem, and then SVM finds the optimal separating hyperplane, which is also the maximal margin hyperplane. The optimization criterion is the width of the margin between the classes, i.e., the empty area around the decision boundary defined by the distance to the nearest training samples. These patterns, called the support vectors, finally define the classification function. The RS method is the combining technique which modifies the training dataset, builds classifiers on these modified training sets, and then combines them into a final decision rule. In this study, it was obtained the best performance combining the classifiers using the "mean rule" [31].

Nanni applied the method to the problem of differential diagnosis of erythemato-squamous diseases. The dataset collected 366 samples presenting 34 attributes: each sample contained 12 clinical features and 22 histopathological features obtained after a biopsy of a patient skin sample. In this study the "age" feature has been removed because in some

samples the "age" feature was missing. The estimated average predictive error rated by 10-fold cross-validation. The obtained results are 97.22%, 97.22%, 97.5%, 98.1%, 97.22%, 97.5%, 97.8% and 98.3% using LSVM, RS, B1_5, B1_10, B1_15, B2_5, B2_10 and B2_15 algorithms. The results improved the average predictive accuracy obtained by a stand-alone SVM or by a RS ensemble of SVMs.

In [32], the Gini-based decision tree method was modified. It normalized the Gini indexes by taking into account information about the splitting status of all attributes. Instead of using the Gini index for attribute selection as usual, Tran [32] used ratios of Gini indexes in order to reduce the biases.

In this study, the learned model was represented by a decision tree. Decision tree classification has been used for predicting medical diagnoses, because it has several advantages in comparison to the most data mining methods. This study used a dermatology dataset which was a database contains 358 tuples with 35 attributes. Skin samples were taken for the evaluation of 22 histopathological features. Tran compared different methods on the problem of erythemato-squamous diseases which is a real problem in dermatology. After preparing data and applying the modified Gini index approach, the result showed that the traditional Gini index approach and the modified Gini index approach had better accuracies in comparison with ID3 and C4.5.

Polat and Gunes [33] improved the classification accuracy in the case of multi-class classification problem. They proposed a novel hybrid classification system based on C4.5 decision tree classifier [34] and one-against-all approach [35] to classify the multi-class problems including dermatology. In this study the features of a patient were represented as a vector of features. C4.5 Decision tree learning is a method in which the learned function is represented by a decision tree. Learned trees can be represented as sets of if-then rules to improve human readability. The aim of C4.5 Decision tree learning is recursively partition data into sub-groups [34]. Consider an M-class problem and N training samples, one-against all approach constructs M binary C4.5 decision tree classifier, each of which separates one class from all the rest. The *i*th C4.5 decision tree classifiers are trained with all the training examples of the *i*th class with positive labels and all the others with negative labels [35].

To test the hybrid classification system based on C4.5 decision tree classifier and one-against-all approach method, Polat and Gunes used the classification accuracy, sensitivity-specificity analysis, and 10-fold cross validation. By applying C4.5 decision tree, they achieved 84.48% classification accuracy for erythemato-squamous diseases using 10-fold cross validation. Also, their proposed method based on C4.5 decision tree classifier and one-against-all approach obtained 96.71% classification accuracy.

Xie and Wang [36] developed a diagnosis model based on support vector machine (SVM) and a novel hybrid feature selection method to diagnose erythemato-squamous diseases.

In this study, the proposed method combined filter method and wrapper method to find out the optimal feature subset from original feature set, where the improved *F*-score was used as evaluation criteria of filter and Sequential forward search (SFS) [37] was used as an evaluation of wrapper.

The original *F*-score is a simple filter technique which measures the discrimination of two sets of real numbers [38]. For more than two sets of real numbers, in this study, *F*-score was improved to measure the discrimination between them. SFS procedure operates in bottom-to-top manner, and starts with an empty set of features. The single best feature of the original feature is determined and added to the set. At each subsequent iteration or step, the best one feature of the remaining original features is added to the set.

Xie and Wang proposed the hybrid feature selection method, named IFSEF. In the process of filter, they calculated the improved *F*-score for each feature, and then sort them in descending order. In the process of wrapper, a subset of the original training set was generated by including the features with top *N F*-scores, where $N = 1, 2, \ldots,$ m and m was the total number of features. Then a grid search was carried out to find the optimized value. This procedure, using SFS, was carried out until all features appeared in the subset. Finally they obtained the SVM diagnosis model which had 98.61% classification accuracy.

Chen and Chang [39] tried to predict and analyze the most commonly-seen skin diseases and their symptoms. In this study, it is focusing on the relationships of the symptoms and disorders to find important factors and rules that affect skin disorders. Three stages are involved in the modeling construction analysis. The first step includes data collection and pre-processing, producing training data and analyzing variables. The second stage involves five experiments, decision tree model, artificial neural network model, a combination of decision tree and artificial neural network model, decision tree with sensitivity analysis and artificial neural network with sensitivity analysis. The last stage presents comparisons and explanations of each various model. The goal of this study is to focus on six major skin diseases and their symptoms. These six types of common skin disorders are Psoriasis, Seborrheic dermatitis, Lichen planus, Pityriasis rosea, Chronic dermatitis and Pityriasis rubra pilaris. This study uses authentic database in a case medical institute as its research subjects, and then compares the different classification results of data mining when applying decision tree and artificial neural network. It was found that the accuracy of the predictive models in the five experiments was as high as 80% and artificial neural network had the best accuracy. It tempted to provide a referential index system for physicians in clinical diagnosis when using the classification technology of artificial intelligence from the results of the five experiments in this study. For instance, by using the factor rules produced from the decision tree model, human errors in the judgment-making process reduces and medical wastes are avoided. Supplementing the diagnosis with the information provided by this study in addition to comprehensive examinations conducted by professional physicians, the final diagnoses were thus made with higher accuracy.

Paja et al. [40] dealt with computer-aided diagnosing and classification of melanocytic skin lesions. This system has also some teaching functions, improves analyzing of datasets based on calculating of values of TDS (Total Dermatoscopy Score) parameter. In order to find a correct diagnose of lesion, the system using five different methods. These methods were classic ABCD rule (based on TDS parameter) [41] [42],

optimized ABCD rule (based on own New TDS parameter) [43], [44], decision tree (based on ID3 algorithm) [45], a model, called by the author genetic dichotomization, based on a linear learning machine with genetic searching for the most important attributes [46] and finally a model based on application of a classificatory from the family of belief networks [47]. Each method was characterized by some different error rate. By using developed internet-based tool, users are enabled to make early, non-invasive diagnosing of melanocytic lesions.

In [48], a contribution in the field of image processing for dermatological of skin lesions was presented. It dealt with an original color image processing technique applied to skin lesion images. Lesion color information gives important parameter about the clinical presentation of the lesion. This study shows that certain normalization technique can be employed to distinguish three types of psoriasis skin diseases. It makes sure only the lesion color information are taken into consideration. It can be used to produce the psoriasis lesion color components without influence by other artifacts such as hair, cloths or other color information. The images can be represented as a color histogram for normalized RGB color space can be used to classify the lesion according to clinical description. Therefore, the classified features can be proposed as an input to a pre-diagnostic system to aid dermatologist in their work.

Recently artificial neural networks have emerged as a useful and effective means of tackling a range of data mining problems. They have been applied in variety domains, including health and medicine. A neural network is a set of connected input/output units in which each connection has a weight associated with it. During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of the input tuples. Neural networks involve long training times and are therefore more suitable for applications where this is feasible. Advantages of neural networks include their high tolerance of noisy data as well as their ability to classify patterns on which they have not been trained. They can be used when you may have little knowledge of the relationships between attributes and classes. They have been successful on a wide range of real-world data, including handwritten character recognition, pathology and laboratory medicine.

Ubeyli [49] illustrated the use of combined neural networks (CNNs) to guide model selection for diagnosis of the erythemato-squamous diseases. The multilayer perceptron neural networks (MLPNN) were used to test the performance of the method on the diagnosis of the erythemato-squamous diseases. CNN models often result in a prediction accuracy that is higher than that of the individual model. This construction is based on a straightforward approach that has been termed stacked generalization [50]. The MLPNN is a nonparametric technique for performing a wide variety of detection and estimation tasks [51]. In this study, the dataset consisted of 34 features with 12 clinical features and 22 histopathological features. In the dataset, the family history feature has a value of 1 if any of the erythemato-squamous diseases have been observed in the family, otherwise it has a value of zero. The age represents the age of the patient and every other feature was given a degree in the range 0-3. In order to develop CNN for the diagnosis of erythemato-squamous diseases, for the first level models six sets of neural networks were used since there were six possible outcomes of the diagnosis of erythemato-squamous diseases. Networks in each set were trained. The predictions of the networks in the first level were combined by a second level neural network. MLPNNs were used at the first level and second level for the implementation of the proposed CNN. In both first level and second level analysis, the Levenberg-Marquardt training algorithm [52] was used. The levenberg-Marquardt algorithm is a least squares estimation algorithm based on the maximum neighborhood idea. The network topology was the MLPNN with a single hidden layer. Each network had 34 input neurons, equal to the number of input feature vectors. The number of hidden neurons was 30 and the number of output was six. The second level neural network was trained to combine the predictions of the first level networks. The second level network had 36 inputs which correspond to the output of the six groups of the first level networks. The targets for the second level network were the same as the targets of the original data. The number of output was six and the number of hidden neurons was chosen to be 30. The classification results denoted that the CNN with 97.77% classification accuracy obtained higher accuracies comparing with the MLPNN with 85.47% classification accuracy.

Brobst et al. [53] presented a Connectionist Expert System (ES) in the domain of dermatology. It will ultimately be used by medical students for instructing the diagnosis of papulosquamous skin diseases. The ES can be used as a tool for diagnosis. It explains the effects of diagnostic parameters in the derivation of variable conclusions. It uses the Back Propagation algorithm to automatically generate a knowledge base from patient cases. Back Propagation is a form of supervised learning for multi-layer nets, also known as the generalized delta rule. The back propagation algorithm looks for the minimum of the error function in weight space using the method of gradient descent. The combination of weights which minimizes the error function is considered to be a solution of the learning problem. It is most often used as training algorithms in current neural network applications. Consider a network with a single real input x and network function F. based on Back Propagation algorithm, the derivative $F'(x)$ is computed in two phases:

Feed-forward: the input x is fed into the network. The primitive functions at the nodes and their derivatives are evaluated at each node. The derivatives are stored.

Back propagation: the constant 1 is fed into output unit and the network is run backwards. Incoming information to a nod is added and the result is multiplied by the value stored in the left part of the unit. The result is transmitted to the left unit. The result collected at the input unit is the derivative of the network function with respect to x [54].

The Back Propagation learning algorithm is a gradient descent algorithm in which weights in the network are modified in order to minimize the overall mean square error between the desired and the actual output values for all output units and for all patterns.

In [53], training data consisted of list of symptoms as input and the expert's diagnosis as the desired output. The knowledge was distributed over the network so that a pattern

of weighted interconnections constituted an implicit specification of decision criteria. This knowledge base was used by the ES to diagnose a patient's disease given a list of symptoms. It becomes possible to automatically generate a knowledge base from patient cases and eases the task of building an expert system.

During the last few years observations show an escalating use of fuzzy set theory in the field of medical diagnostics. Obviously, fuzzy set theory responds effectively to the non-statistical uncertainty, which is circumscribed in problems of the medical domain. Fuzzy systems usually generate human interpretable rules which take the form of IF-THEN statements. Such rules which correspond to the knowledge granules of the diagnostic systems can be comprehended by human operators.

Lekkas and Mikhailov [55] reviewed a methodology for evolving fuzzy classification which allows data to be processed in online mode by recursively modifying a fuzzy rule base on a per-sample basis from data streams. They presented a study of semi-supervised evolving fuzzy classification on the diagnostics for two well known medical problems. The first one was Pima Indians diabetes (PID) and the second one was about the classification of six dermatological diseases (DERM) which are considered as erythemato-squamous diseases. The DERM dataset contains 366 data samples, 8 of which have been removed for they contain missing values. Each sample consists of 34 numerical features, 12 of which are of clinical nature and 22 of histopathological nature.

The vast majority of existing methodologies for fuzzy medical diagnostics require the data record to be processed in offline mode, as a batch. Unfortunately this allows only a snapshot of the actual domain to be analyzed. Should new data records become available they require cost sensitive calculations due to the fact that re-learning is an iterative procedure. Evolving fuzzy classification systems (EFCS), such as eClass operate by self-developing their fuzzy rule base in one pass on a per-sample basis. Thus, their fuzzy rule base is not fixed as with the offline counterparts but instead it can adapt to the information brought by data samples which arrive from massive data streams sequentially. The adaptation process does not require re-learning, because it is based on recursive calculations. However, eClass is data order dependent as different orders of the data result into different rule bases. Nonetheless, in [55], they showed that models of eClass can be improved by arranging the order of the incoming data using a simple optimization strategy.

Evolving fuzzy rule-based classifier can start either with an empty rule-base or with some pre-specified set of fuzzy rules. Each new data sample that has been read can be used to upgrade or modify the rule base if the label is also provided. If the label is not provided, the existing fuzzy rule base will generate the predicted class. eClass can work in any combination of these two modes. An important specific of eClass is that not only the number of fuzzy rules but the number of classes may also be evolving and does not need to be pre-fixed [56]. The process of learning the antecedent (if-part) of the fuzzy rules is based on an online fuzzy clustering method called eClustering [56], and the process of learning the consequent of a first order Takagi-Sugeno-Kang (TSK) fuzzy rule is realized by approximating the nonlinear n-dimensional input using multiple locally linear sub models, one per class *c* of the C-class problem. The inference process complements the one of learning, since without the former the latter is inessential. The system is aimed to learn in order to make predictions, which are based on the product of its learning, its knowledge.

In this study, eClass algorithm was applied on DERM dataset and four different configurations had been used per dataset. The first configuration indicated that the data will be processed in their original order. The rest three configurations utilized the buffering technique in order to realize different data orders. The results of applying the four aforementioned configurations on the DERM dataset showed that the buffering technique had a positive impact on the performance of the two models. It obtained the accuracy of 97.55%.

Ubeyli and Guler [57] proposed an approach based on adaptive neuro-fuzzy inference system (ANFIS) for detection of erythemato-squamous diseases. Neuro-fuzzy systems are fuzzy systems which use artificial neural networks theory in order to determine their properties by processing data samples. A specific approach in neuro-fuzzy development is the adaptive neuro-fuzzy inference system, which has shown significant results in modeling nonlinear functions. In [57], the six ANFIS classifiers were used to detect the six erythmato-sequamous diseases when the inputs were 34 features defining six disease indications. The performance of the ANFIS model was evaluated in terms of training performances and classification accuracies. The proposed approach obtained 95.5% classification accuracy which was higher than that of the stand-alone neural network model. The total classification accuracy of the stand-alone neural network was 85.5%. The results showed that the proposed ANFIS model has some potential in detecting the erythemato-squamous diseases.

Genetic programming (GP) is a stochastic population-based search method devised in 1992 by John R. Koza. GP is an extension of genetic algorithm. It is a general search method that uses analogies from natural selection and evolution. GP encodes multi-potential solutions for specific problems as a population of programs or functions [58].

Bojarczuk et al. [59] used classification for diagnosing certain pathologies. In order to discover the rules, a constrained-syntax GP algorithm was developed which was based on some concepts of data mining, particularly with emphasis on discovering the comprehensible knowledge. Three experiments were done using databases of medical domain. One of them was in dermatological domain. All attributes had values mapped in the range [0...3], except age (integer, in years) and family history (1, 0). The database was randomly divided into five partitions. Then a well-known 5-fold cross validation procedure was performed. GP was run once for each class. Once all runs of GP for a given dataset was completed, all the rules found by GP in that experiment were grouped into a rule set. The quality of that rule set was evaluated according to the predictive accuracy of the rule set and the comprehensibility of the rules. The results showed that the proposed GP obtained 96.64% classification accuracy rate considerably better than C4.5 a very well known advanced decision-tree algorithm often used in machine learning and

data mining. C4.5 obtained 89.12% classification accuracy rate, Overall, the rule set discovered by the GP was simpler (shorter) than the rule sets discovered by C4.5.

D. Clustering

As opposed to the classification task, when we carry out cluster analysis, the class labels of data objects or instances are not present in the training set, and hence the type of learning that needs to occur is of an unsupervised type. The main aim of the clustering methods is to group the data objects according to some measure of similarity, so that the data objects in one group or cluster are highly similar to one another while being very different from the data objects belonging to the remaining cluster. In other words, the aim is to maximize the intra-class similarity and minimize the inter-class similarity [1]. In the health care profession, this type of problem is interesting to discover information about a drug, a treatment or a disease [22].

Ubeyli and Dogdu [60] presented an approach based on the implementation of *k*-means clustering [61] for automated detection of erythemato-squamous diseases. *k*-Means is one of simplest algorithms that can be used to partition a number of data points into a given number of clusters. It expects that the number of clusters is given as input to the algorithm. "*k"* in the name *"k*-means" stands for the number of clusters [61]. *k*-means algorithm assumes that each data point has a single comparable numeric value. Otherwise, when data points have multi-attribute values, distances between points are calculated using Euclidian distance. In this study, the algorithm's task is to assign the data points to one of the five clusters. They experimented with five class, therefore they removed the data for the patients who have the diseases of the sixth class and the algorithm was used to detect five disease indications were used. They obtained 94.22% prediction accuracy of the *k*-means clustering. So they concluded that the proposed *k*-means clustering can be used in detecting erythemato-squamous diseases by taking into account the misclassification rates.

Tasoulis and Doukas [62] proposed a novel clustering technique for the characterization and categorization of pigmented skin lesion in the dermatological image. To segment a cutaneous image, the skin lesion to be assessed must be separated from the healthy skin. Mainly, region-based segmentation methods are being used for this purpose [63] [64]. The proposed method by Tasoulis and Doukas used Principal Component Analysis. They tried to produce high quality retrieval of melanoma samples in skin lesion images. The results have proved the superiority of this method against traditional ones.

In [65], a spatial data mining method to skin lesion image and its integration with a segmentation method was adopted to identify significant color regions in an image. The dataset that the algorithm was applied to was a set of split regions generated in the process of splitting and their objective was to cluster those split regions to form segments/objects of interest. In the preprocessing phase, each image was smoothed before it goes to segmentation. The purpose for preprocessing was to remove noise such as reflection. After that a region is split into four sub-regions if it fails to pass homogeneity test. The process of splitting goes recursively until all the regions are found homogenous or are too small to be split. In this case, a clustering technique, DBSCAN (Density Based Spatial Clustering of Applications with Noise) [66] was combined with image processing techniques. DBSCAN is designed to discover clusters and noise in a spatial database. It is used to segment skin lesion images, identify sub-region inside lesions and extract color features. The resulting regions were compared with human reception via Kappa statistical test. Evaluation of the results indicates that the method approximates human judgment well and can be used as an automatic tool for mining skin lesion images.

Mohammadi and Duzgun [67] concentrated on the spatial distribution of spotted pattern on the surface of the skin. The point pattern of the spot's spatial distribution was tested if it was random, regular, or clustered. The exploration and modeling of the point pattern could help developing appropriate treatment and amount of the medicine to be applied, especially making advises through e-media. To visualize data, quadrat density method [68] was used. A quadrat is a frame of any shape. The goal of the quadrat method is to estimate the population density of each species in a given community. Population density is the number of individuals of each species per unit area. Small square areas, called quadrats, are randomly selected to avoid choosing unrepresentative samples. Once the population densities for all quadrates are determined, the population size within the large area can be estimated. What is gained from quadrat method helped to get a general idea of distribution and density of point pattern on the skin surface. Then kernel estimation was used to obtain a smooth estimate of probability density from the observed area and do nearest neighbor distance tests for Complete Spatial Randomness (CSR). Moreover the CSR test was applied by using K-function and L-function for the some points to see if the observed pattern was clustered, regular or random. The most infected areas on the skin (indicating clustering) were exposed to the pendant ingredients and polluted air which normally is not covered by cloths. CSR tests indicated clustering in points marks.

Table I. Some studies which have worked on skin diseases mining

Skin Disease/ DM approach	Skin Cancer	6 skin diseases*	Melanocytic skin lesions	Erythmeato-squamous diseases	Skin lesion images	Papulosquamous skin diseases
Association	Yes	No	No	No	No	No
classification	No	Yes	Yes	Yes	Yes	No
Clustering	No	No	No	Yes	Yes	No

* These six types of common skin disorders are Psoriasis, Seborrheic dermatitis, Lichen planus, Pityriasis rosea, Chronic dermatitis and Pityriasis rubra pilaris

V. DISCUSSION

The summary of application of different data mining methods in skin diseases has been shown in the Table I. As it can be seen in the Table I, the classification method has been used in many studies. In [39], [40] and [32] , the decision tree mining has obtained relatively good results. In [32] the Gini-based decision tree has been modified. The results show that their method performs better in comparison to ID3 and C4.5. Nevertheless, in [39] is expressed that artificial neural network is a better method for extracting rules. In the case of skin lesion images, in addition to classification, [48], clustering methods, [65], have been used. And observations show that, the proposed method by [65] can be replaced human judgments.

As it is mentioned in Table I, association mining has been used in skin cancers studies. In [17] and [19] which have applied association mining as the method for extracting rules, Apriori algorithm is known as the best method. The summary of usage of different association rule mining algorithms in dermatology has been shown in Table III.

The relations between symptoms and disorders are really important. In most of studies, it is tried to discover the relationship between symptoms and diseases. Diagnosis of erythemato-squamous diseases is a real and difficult problem in dermatology.

For comparison purpose, Table II gives the classification accuracy of the different methods applied for diagnosis of erythemato-squamous diseases in the last decade.

Bojarczuk [59], Tan [32] and Polat and Gunes [33] have proposed the different methods, but they all compared their method with C4.5 and the results show that their approaches had better accuracies in comparison with C4.5. In the case of Neural networks, Ubeyli [49] used the combined neural network models. The CNN with 97.77% classification accuracy obtained higher accuracies comparing with the Ubeyli and Guler's method [57] which was based on adaptive neuro-fuzzy inference system and had 95.5% classification accuracy. Lekkas and Mikhailov [55] used evolving fuzzy classification. This model is as much as 97.55% accurate on the differential diagnosis of the 6 erythemato-squamous diseases. Except [55], all the aforementioned algorithms are designed for offline use and hence they cannot manage online data efficiently form a computational point of view.

In the use of clustering techniques, Ubeyli and Dogdu [60] proposed the method based on *k*-mean clustering which had 94.22% classification accuracy.

Most studies about diagnosis of erythemato-squamous diseases are in the classification section. From Table II, we can see that there are various classification methods for diagnosis of erythmato-squamous disease from 1998 up to now. Among these studies, Guvenir and Emeksiz [25] had the highest classification accuracy, 99.2%, on the differential diagnosis of erythemato-squamous diseases using voting feature intervals-5 and 10-fold cross validation.

Nanni [28] and Xie and Wang [36] both used SVM and feature selection methods. The method proposed by [36] achieved 98.61% classification accuracy and its accuracy rates and the model proposed by Nanni [28] are very close.

Table II. Classification accuracies obtained with different classifiers for diagnosis of erythemato-squamous diseases.

Author	**Year**	**Method**	**Classification accuracy**
Guvenir et al. [23]	1998	VFI5	96.2%
Guvenir and Emeksiz [25]	2000	Nearest neighbor classifier Naïve Bayesian classifier VFI5	99.2%
Bojarczuk [59]	2001	A constrained-syntax genetic programming C4.5	96.64% 89.12%
Ubeyli and Guler [57]	2005	ANFIS	95.5%
Nani [28]	2006	LSVM RS B1_5 B1_10 B1_15 B2_5 B2_10 B2_15	97.22% 97.22% 97.5% 98.1% 97.22% 97.5% 97.8% 98.3%
Polat and Gunes [33]	2009	C4.5 and one-against-all	96.71%
Ubeyli [49]	2009	CNN	97.77%
Ubeyli and Dogdu [60]	2010	K-mean clustering	94.22%
Lekkas and Mikhailov [55]	2010	Evolving fuzzy classification	97.55%
Xie and Wang [36]	2011	IFSFS and SVM	98.61%

VI. CONCLUSION

In this paper, data mining and statistical techniques used in medicine have been discussed. As knowledge underlying health and illness is improving, the available data are becoming more numerous and complex, so many demands are arrived to process these data. Medical data mining can help to prepare some methods for diagnosis, prognosis, decision making, etc. In this study, we have summarized some uses of data mining techniques in medical domain. We have focused on the application of data mining in skin diseases in the past two decades. Diagnosis of erythemato-squamous diseases is a difficult problem in dermatology. There are several studies which have worked on it using dermatology datasets. In this paper we have summarized some of them and reported their results.

Table III. Association rule mining algorithms used for mining of skin diseases.

Author	**Association rule mining algorithm**	**usage**
Wang and Chung [17]	SETM algorithm	Finding association rules between different features of skin cancer images.
Tickle and Nahar [19]	Apriori algorithm Predictive appriori algorithm Tertius algorithm	Discovering most significant risk factors for skin cancer.

REFRENCES

[1] Jiawei Han and Micheline Kamber, *Data Mining Concepts and Techniques*. San Francisco, CA: Elsevier Inc, 2006.

[2] U. M., Piatetsky-Shapiro, G. & Smyth, P. & Uthurusamy, R. Fayyad, "From Data Mining to Knowledge Discovery: An Overview," in *Advances in Knowledge Discovery and Data Mining* , 1996a, pp. 1-36.

[3] S.-C. Liao & M. Embrechts I. -N. Lee, "Data mining techniques applied to medical information ," *Med. Inform* , pp. 81-102, 2000.

[4] E., Donald, "Introduction to Data Mining for Medical Informatics ," *Clin Lab Med*, pp. 9-35, 2008.

[5] R., Zhang, Y., Katta, "Medical Data Mining," *Data Mining and Knowledge Discovery*, pp. 305-308, 2002.

[6] Irene M. Mullins et al., "Data mining and clinical data repositories: Insights from a 667,000 patient data set," *Computers in Biology and Medicine*, vol. 36, pp. 1351-1377, 2006.

[7] J. C., Lobach, D. F., Goodwin, L. K., Hales, J. W., Hage, M. L., EdwardHammond, W. Parther, "Medical Data Mining: Knowledge Discovery in a Clinical Data Warehouse," 1997.

[8] Dan Dalan, "Clinical data mining and research in the allergy office," *Current Opinion in Allergy & Clinical Immunology*, vol. 10, no. 3, pp. 171-177, June 2010.

[9] Y., Mahajani, G., Aslandogan, "Evidence Combination in Medical Data Mining ," in *Information Technology: Coding and Computing (ITCC)*, 2004.

[10] K.J. and G.W. Moore Cios, "Uniqueness of medical data mining ," *Artificial Intelligence in Medicine*, pp. 1-24, 2002.

[11] Y. liu and Y. F. Zheng, "FS_SFS: A novel feature selection method for support vector machines," *Pattern Recognition*, vol. 39, pp. 1333–1345, 2006.

[12] H. Liu and R. Setiono, "A probabilistic approach to feature selection - A Filter Solution," pp. 319-327, 1996.

[13] L. Talavera, "An evaluation of filter and wrapper methods for feature selection in categorical clustering.," in *Proceedings of 6th international symposium on intelligent data analysis*, Madrid, Spain, 2005, pp. 440-445.

[14] G., Rayward-Smith, V.J., Sonksen, P.H., Carey, S., Weng, C., Richards, "Data mining for indicators of early mortality in a database of clinical records," *Artificial Intelligence in Medicine*, pp. 215-231, 2001.

[15] R., Imielinski, T., Swami, A.N., Agrawal, "Mining Association Rules between Sets of Items in Large Databases," in *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data* , Washington, D.C., 1993, pp. 207-216.

[16] Li Jiuyong, Ada Wai-chee Fu, and Paul Fahey, "Efficient Discovery of Risk Patterns in Medical Data," *Current Opinion in Allergy and Clinical Immunology* , vol. 10, no. 3, pp. 171-177, 2010.

[17] S.M., Wang , Q., Chung, "Content-based Retrieval and Data Mining of a Skin Cancer Image Database," in *Proccedings of International conference on Information Technology Coding and Computing*, 2001, pp. 611-615.

[18] M., Swami, A., Houtsma, "Set-Oriented Mining for Association Rules in Relational Databases," in *Proceedings of International Conference on Data Engineering*, 1995, pp. 25-33.

[19] J., Tickle, K.S., Nahar, "Significant Cancer Risk Factor Extraction: An Association Rule Discovery Approach," in *Proceedings of International Wrokshop on Data Mining and Artificial Intelligence*, Khulna, 2008, pp. 108-114.

[20] T. Scheffer, "Finding Association Rules that Trade Support Optimally Against Confidence.," in *Proceedings of the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases(PKDD'01)*, Freiburg, Germany, 2001, pp. 424-435.

[21] P.A. Flach and N. Lachiche, "Confirmation-guided discovery of first-order rules with Tertius," *Kluwe Academic Publishers*, pp. 61-95, 2001.

[22] A.L., Chen, H., Hubbard, S.M., Schatz, B.R., NG, T., Seweel, R.R., Tolle, K.M., Houston, "Medical Data Mining on the Internet: Research on a Cancer Information System," *Artificial Intelligence Review*, pp. 437-466, 2000.

[23] H. A. Govenir, G. Demiroz, and N. Ilter, "Learning differential diagnosis of Eryhemato-Squamous diseases using voting feature intervals," *Aritificial Intelligence in Medicine*, vol. 13, pp. 147-165, 1998.

[24] G. Demiroz and H. A. Guvenir, "Classification by Voting Feature Intervals," in *Proceedings of Ninth European Conference on Machine Learning(ECML-97)*, 1997, pp. 85-92.

[25] H.A. Guvenir and N. Emeksiz, "An expert system for the differential diagnosis of erythemato-squamous diseases," *Expert Systems with Applications*, vol. 18, pp. 43–49, 2000.

[26] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learining algorithms ," *Machine Learining* , vol. 6, pp. 37-66, 1991.

[27] R. O. Duda and P. E. Hart, *Pattern classification and sence analysis*. New York: Wiley, 1973.

[28] L. Nanni, "An ensemble of classifiers for the diagnosis of erythemato-squamous diseases," *Neurocomputing*, vol. 69, pp. 842-845, 2006.

[29] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*. New York: Wiley, 2000.

[30] T.K. Ho, "The random subspace method for constructing decision forests," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 8, pp. 832-844, 1998.

[31] J. Kittler, M. Hatef, R. Duin, and J. Matas, "On combining classifiers," *IEEE transactions on Pattern Analysis and Machine Intelligence* , vol. 20, no. 3, pp. 226-239, 1998.

[32] Q., Tran, "Mining Medical Databases with Modified Gini Index Classification," in *Fifth International Conference on Information Technology: New Generations*, 2008, pp. 195-200.

[33] K. Polat and S. Gunes, "A novel hybrid intelligent method based on C4.5 decision tree classifier and one-against-all approach for multi-class classification problems," *Expert Systems with Applications*, vol. 36, no. 2, pp. 1587-1592, 2009.

[34] M. T. Mitchell, *Machine Learning*. Singapore: McGraw-Hill, 1997.

[35] L. Yi and Y. F. Zheng, "One-against-all multi-class SVM classification using reliability measures," in *Proceedings of the IEEE international joint conference on neural networks IJCNN '05*, 2005, pp. 849-854.

[36] J. Xie and Ch. Wang, "Using support vector machines with a novel hybrid feature selection method for diagnosis of erythemato-squamous diseases," *Expert Systems with Applications*, vol. 38, no. 5, pp. 5809–5815, 2011.

[37] S. Gunal, O. N. Gerek, D. G. Ece, and R. Edizkan, "The search for optimal feature set in power quality even classification. ," *Expert Systems with Applications*, vol. 36, pp. 10266-10273, 2009.

[38] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert Systems with Applications*, vol. 36, pp. 3240-3247, 2009.

[39] Ch., Chen , Ch., Chang, "Applying decision tree and neural network to increase quality of dermatologic diagnosis," *Expert Systems with Applications*, pp. 4035-4041, 2009.

[40] T., Paja, W., Piatek, L., Wrzesien, M., Mroczek, "Classification and Synthesis of Medical Images in the Domain of Melanocytic Skin Lesions," in *HSI*, Krakow, Poland, 2008, pp. 705-709.

[41] O., Stolz, W., Bilek, P. Braun-Falco and T., Landthaler, M., Merkle, "Das Dermatoskop," in *Eine Vereinfachung der Auflichtmikroskopie von pigmentierten* , Hautartzt, 1990, pp. 131-136.

[42] W., Harms, H., Aus, H.M., Abmayr, W., Braun-Falco, O., Stolz, "Marcoscopic diagnosis of melanocytic lesions using color and texture image analysis," in *J. Invest. Dermatol*, 1990, pp. 491-497.

[43] A., Bajcar, S., Brown, F.M., Grzymala-Busse, J.W., Hippe, Z.S., Alvarez, "Optimization of the ABCD Formula Used for Melanoma Diagnosis," in *Advance In Computing (Intelligent Inoformation Systems and Web Mining)*, Heidelberg, 2003, pp. 233-240.

[44] J.W., Bajcar , S., Grzymala-Busse, W.J., Hippe, Z.S., Grzymala-Busse, "Data Mining Analysis of the ABCD Formula Used for Diagnosis of Melanoma ," in *Concurrency Specification and Progmamming*, Czarna (Poland), 2003, pp. 25-27.

[45] J.W., Hippe, Z.S., Knap, M., Mroczek, T., Grzymala-Busse, "Nowy Algorytm generowania drzew decyzji," in *Biocybernetyka i inzynieria biomedyczna*, Gdansk, 2003, pp. 257-262.

[46] Z.S., Wrzesien, M., Hippe, "Some Problems of Uncertainty of Data after the Transfer from Multi-category to Dichotomous Problem Space," in *Methods of Artificial Intelligence*, Gliwice, 2002, pp. 185-189.

[47] Z.S., Mroczek, T., Hippe, "Melanoma classification and prediction using belief networks," in *Computer Recognition Systems KOSYR, Univ. of Technology Edit*, Wroclow, 2003, pp. 337-342.

[48] R., Hashim, H., Taib, M.N., Jailani, "Normalization Techniques for Psoriasis Skin Lesion Analysis," in *New Techniques in Pharmaceutical and Biomedical Recearch*, Kuala Lumpur, Malaysia, 2005, pp. 151-153.

[49] E. D. Ubeyli, "Combined neural networks for diagnosis of erythemato-squamous diseases," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5107–5112, 2009.

[50] J. V. Hansen and R. Nelson, "Data mining of time series using stacked generalizaers.," *Neurocomputing*, vol. 43, pp. 173-184, 2002.

[51] I. A. Basheer and M. Hajmeer, "Artificial neural networks: fundamentals, computing, design, and application," *Journal of Microbiological Methods*, vol. 43, no. 1, pp. 3-31, 2000.

[52] Kenneth Levenberg, "A Method for the Solution of Certain Non-Linear Problems in Least Squares," *The Quarterly of Applied Mathematics*, vol. 2, pp. 164-168, 1944.

[53] Y., Brobst, R.W., Bergstresser, P.R., Peterson, L., Yoon, "Automatic Generation of a Knowledge-Base for a Dermatology Expert System," in *Third Annual IEEE Symposium on Computer-Based Medical Systems*, 1990, pp. 306-312.

[54] Raul Rojas, *Neural networks: a systematic intorduction*. Berlin: Springer-Verlag, 1996.

[55] Stavros Lekkas and Ludmil Mikhailov, "Evolving fuzzy medical diagnosis of Pima Indians diabetes and of dermatologica diseases," *Artificial Intelligence in Medicine*, vol. 50, pp. 117-126, 2010.

[56] P. Angelov, X. Zhou, and F. Klawonn, "Evolving fuzzy rule-based classifier," in *In Proceeding of the IEEE symposium on computational intelligence applications in image and signal processing*, 2007, pp. 220-225.

[57] E. D. Ubeyli and I. Guler, "Automatic detection of erythemato squmous diseases using adaptive neuro-fuzzy inference systems," *Computer in Bialogy and Medicine*, vol. 35, pp. 421-433, 2005.

[58] J.R. Koza, *Genetic Programming II: Automatic Discovery of Reusable Programs.*: MIT Press, 1994.

[59] C. C., Lopes, H. S., Freitas, A. A., Bojarczuk, "Data Mining with Constrained-Syntax Genetic Programming: Applications in Medical Data Set," in *Data Analysis in Medicine and Pharmacology (IDAMAP-2001), a Workshop at Medinfo-2001*, London, UK, 2001.

[60] E. D. Ubeyli and E. Dogdu, "Automatic Detection of Erythemato-Squamous Diseases Using k-Means Clustering," *Journal of Medical Systems*, vol. 34, pp. 179-184, 2010.

[61] H. D. Margaret, *Data Mining: Introductory and Advanced Topics.*: Prentic-Hall, 2003.

[62] S. K. Tasoulis and C. N. Doukas, "Classification of Dermatological Images Using Advanced Clustering Techniques," in *32nd Annual International Conference of the IEEE EMBS*, Buenos Aires, Argentina, 2010, pp. 6721-6742.

[63] D. H. Chung and G. Sapiro, "Segmenting skin lesions with partial-differenctial-equations-based image processing algorithms," *IEEE Transactions on Medical Imaging*, vol. 19, no. 2, pp. 763-767, 2000.

[64] Z. Zhang, W. Steocker, and R. Moss, "Border detection on digitized skin tumor image," *IEEE transactions on Medical Imaging*, vol. 19, no. 11, pp. 1128-1143, 2000.

[65] W., Aslandogan, Y., Guo, "Mining Skin Lesion Images with Spatial Data Mining Methods," Arlington, 2003.

[66] M. Ester, H.P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering cluster in large spatial databases," in *proccedings of International Conference of Knowledge Discovery and Data Mining (KDD'96)*, Portland, OR., 1996, pp. 226-231.

[67] E., Duzgun, S., Mohammadi, "Statistical Data Analysis and Modeling of Skin Diseases," , 2008.

[68] R. Pound and F.E. Clements, "A method of determining the abundance of secondary species," in *Botanical Studies, II*, Minnesota, 1998, pp. 19-24.

[69] S.M., Wang , Q., Chung, "Content-based Retrieval and Data Mining of a Skin Cancer Image Database," in *in Proc. of Int. Conf. on Information Technology Coding and Computing* , 2001, pp. 611-615.

[70] M., Swami, A., Houtsma, "Set-Oriented Mining for Association Rules in Relational Databases," in *Proc. of Int'l Conference on Data Engineering*, 1995, pp. 25-33.

[71] Charu Gupta, "IMPLEMENTATION OF BACK PROPAGATION ALGORITHM (of neural networks) IN VHDL," Department Of Electronics and Communication Engineering THAPAR INSTITUTE OF ENGINEERING & TECHNOLOGY,(Deemed University), Patiala, India, Master thesis 2006.