Chapter One

Introduction

1.1 Introduction

Dermatology is a branch of medical science that deals with skin diseases, In dermatology there are a lot of diseases which shows similar features in appearance and symptoms.(MadhuraRambhajani1 et al., 2015)

All the diseases will be classified at this study share the same clinical feature of erythema and scaling with very little difference, this similarity in symptoms makes diagnosis more difficult, there are two kinds of attributes in the medical data:

1- Clinical Attributes :

Clinical feature is medical sign which indicate some medical fact or characteristic that may be detected by a patient or anyone, especially the doctor, before or during a physical examination of a patient. For example, whereas a tingling paresthesia is a symptom (only the person experiencing it can directly observe their own tingling feeling), erythema is a sign (anyone can confirm that the skin is redder than usual).

Some signs may have no meaning to the patient, and may even go unnoticed, but may be meaningful and significant to the doctor

2- Histopathological attributes

Histopathology is the microscopic examination of biological tissues to observe the appearance of cells and tissues in very fine detail.

1.2 Problem statement

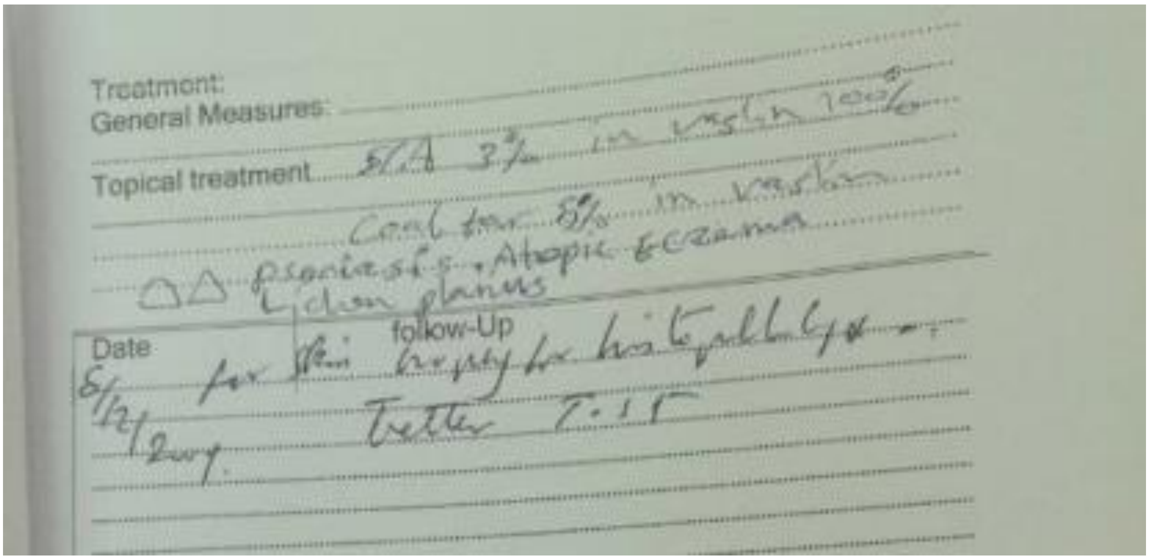
A huge amount of data is being constructed in Sudanese hospitals, but unfortunately doctors does not gain benefit from it in an optimal manner. Such data include wealth of knowledge but are buried and not utilized.

Medical diagnosis is an important task but sometimes may be little complicated. Diagnosing of one patient can vary if the patient is examined by different doctors or even by the same doctors at different times. Automated medical diagnosis support the doctors to predict the right disease with less time(K. et al., 2014).

Sometimes the diagnosis of skin diseases becomes very complex due to the similarity of their characteristics. Which is why doctors may not be able to diagnose the disease in time, thus it may turn into skin cancer(MadhuraRambhajani1 et al., 2015).

The problem was evident during the collection of data from patient's medical reports, doctors may sometimes be unable to accurately identify the real disease and give several probabilities of the disease, as a result the detection of the real disease is delayed and may turn into a more serious disease, figure 1.1 shows differential diagnosis of a patient depending on the symptoms, the doctor suggested that the disease may be either Psoriasis, lichen planusor Atopic Eczema.

Another difficulty for the differential diagnosis is that a disease may show the features of another disease at the beginning stage and may have the characteristic features at the following stages.



Treatment:
General Measures:
Topical treatment S/A 3% in vaselin 100%
Coal tar 8% in vaselin
ΔΔ psoriasis, Atopic Eczema
Lichen planus

Date
8/12/2007

follow-Up
for skin biopsy for histopathology
Better [illegible]

Figure1.1 Differential Diagnosis

1.3 Research Significance

The correct medical diagnosis is very important process, but sometimes it is difficult for the doctor to know the exact disease. The medical programs support the doctor in diagnosis. In order to avoid delaying diagnosis and treatment, we use data mining techniques to build classification models for predicting some dermatology diseases which have the similar symptoms, accordingly increasing the quickness of diagnosis process to prevent worsen of the disease.

1.4 Hypotheses

Using data mining techniques to build a model diagnosing some Dermatology diseases in spite of absence of histopathological features.

1.5 Objectives

1- Build a prediction model for diagnosis some dermatology diseases.

2- Build a real dermatology dataset from real Sudanese patients reports

1.6 Research Scope

The scope of this research is as follows:

• The classification models diagnosis the four following dermatology diseases:

1- Psoriasis

2- seboreic dermatitis

3- lichen planus

4- cronic dermatitis

• The used dataset is a real Sudanese data collected from Omdurman Military Hospital department of Dermatology, this data set just contain the clinical symptoms for the above diseases, doctors don't write the histopathological attributes in the patient medical reports.

1.7 Research Organization

This research is divided into five chapters as follows:

Chapter one give general introduction about the research, it consist the problem statement, Research Significance, Hypotheses**,** objectives and scope. Chapter two is about Literature review and related work. Chapter three include the Methodology and techniques. Chapter four contain data set description, preprocessing, Implementation phase, Result Interpretation and Analysis. Chapter Five provide Conclusion and recommendation.

Chapter Two

Literature Review and Related Work

2.1 Introduction

This chapter consists of two sections, the first section represent a brief overview about the datamining and the second section is about related works.

2.2 Knowledge Discovery from Data (KDD)

"Knowledge Discovery from Data is a process that extract knowledge from large amounts of data, turning such data into useful information, Knowledge discovery as a process is consist of the following steps sequentially"(Han and Kamber, 2006):

1. **Data cleaning** (to remove noise and inconsistent data)
2. **Data integration** (where multiple data sources may be combined)
3. **Data selection** (where data relevant to the analysis task are retrieved from the database)
4. **Data Transformation:** The data even after cleaning are not ready for mining as we need to transform them into forms appropriate for mining. The techniques used to accomplish this are smoothing, aggregation, normalization etc.
5. **Data mining** (an essential process where intelligent methods are applied in order to extract data patterns)
6. **Pattern evaluation** (to identify the truly interesting patterns representing knowledge based on some interestingness measures
7. **Knowledge presentation** (where visualization and knowledge representation techniques are used to present the mined knowledge to the user.

2.2.1 Data mining functionality

Data mining tasks can be classified into two categories (Han and Kamber):

1- **The predictive model** which make prediction about unknown or missing data values by using the known values. eg. Classification, Regression, Prediction, Time series analysis etc .

2- **The descriptive model** which identifies the patterns or relationships in data and explores the properties of the data examined eg. Clustering, Association rule, Summarization, Sequence discovery etc.

2.3 Data mining in medical domain

In recent years, data mining is used in different medical domain such as: genetics, medicine and biomedical. Data mining tools increase the performance of diagnostics, prevention and treatment of the diseases.

The aim of using Data mining technique in medical domain is to support hospitals, clinics, physicians, and patients by early detection of life threatening diseases, the patient's data might contain potential facts or details about the possibility of diseases that can occur in a patient. This knowledge can be a great help for better diagnosis and treatment for future medical cases(Thorat1 and Kute2, 2014).

Data mining has great ability to explore the hidden patterns in medical data sets and use these patterns for clinical diagnosis, just the available raw medical data need to be collected in an organized form. Then data mining technology applying on such data to discover the hidden patterns. Data mining and statistics both strive towards discovering patterns and structures in data. Statistics deals with heterogeneous numbers only, whereas data mining deals with heterogeneous fields(Wasan1 et al., 2006).

The biggest challenge facing data mining in the health sector is the quality of its data. It is very difficult to obtain accurate and complete health care data. Health data are complex and heterogeneous in nature because they are collected from various sources such as medical reports for the laboratory, from a patient's discussion or from a physician's review. Without accurate and correct data we will never achieve useful results. Therefore, it is necessary to gain such data in order to extract useful information for an effective decisions(Thorat1 and Kute2, 2014).

2.4 Related work

This section presents four previous studies related to the using of data mining techniques in diagnosis some dermatology diseases, including the used techniques and the obtained results.

2.4.1 Using Naïve Bayesian Classification

Manjusha K.et al. proposed a system to obtain data patterns with the help of Naïve Bayesian theorem. They gathered the medical data from tertiary health care centers which treating people from various areas of Kottayam and Alappuzha, Kerala, India.Their main goal of the research is to analyze the data and to decide whether it is suitable to be analyzed with the use of the data mining methods or not, then predict the probability of occurring eight dermatological diseases which are Scarlet fever, fifth disease chicken box ,scarlet fever, Rubella , measles, enterovirus and subitum. The medical data set contains 230 instance and has 21 medical attributes. They built a model to predict the occurring of these diseases using the Naïve

Bayesian algorithm, their results show that the chances were more predictable for scarlet fever disease and very little probability for Kawasaki and chickenpox diseases (K. et al., 2014).

For future consideration they indicate that this work can be extended with other techniques, also can predict other diseases.

2.4.2 Classification using Bayes net and Best First Search

Madhura Rambhajani1 et al. used the dermatology dataset which is available in UCI repository site, they classify it using the Bayes Net with Best First search. The dataset contain 365 instance and 35 attributes, 34 attributes are considered as input and the 35th attributes is considered as target (class), this study aimed to classify six dermatology diseases, the best result obtained after eliminating fifteen features , the accuracy was %99.31 (MadhuraRambhajani1 et al., 2015).

2.4.3 Naive Bayes and J48

Manjusha K. K et al. Gathered medical dataset from the southern part of Kerala, India. It contain 230 instances with 22 attributes. They applied two data mining classification algorithms Naive Bayes' (NB) and J48 for data analysis. To improve the efficiency of the process they used 10 folds cross validation.

Their result showed that Naive Bayes produced less precision and true Positive rate as compared with J48 algorithm. J48 is more efficient in all parameter like TP-rate, FP-rate, Precision, Recall and ROC area, when they compare the results base on time they found that Naive Bayes classification requires only 0.01seconds and J48 requires 0.03seconds(K et al., 2015).

2.4.4 Artificial Neural Network and Support Vector Machine

Krupal S. Parikh1 et. al. built two predictive models using Artificial Neural Network and Support Vector Machine. They got the database from Department of Skin & V.D., Shrikrishna Hospital, Karamsad, Gujarat, India. The dataset includes 470 instances and 47 features.

To evaluate the performance of the classifiers, they divided the dataset into 80-20% and 70-30% partitions, i.e., 80% and 70% data for training and 20% and 30% for testing respectively, their result which obtained from both classifiers **ANN** and **SVM** showed that the performance of ANN is more accurate for two hidden layers than SVM (Parikh et al., 2015).
The following table (2.1) shows the analysis of related works:

Study	Techniques	Results	Open Issues
(Sistla Karthik*, 2017)	Naïve Bayesian	the chances were more predictable for scarlet fever disease and very little probability for Kawasaki and chickenpox diseases	Extended with other techniques and predict other diseases
(MadhuraRambhajani1 et al., 2015)	Bayes Net with Best First search	the best result obtained after eliminating fifteen features , the accuracy was %99.31.	Ignore the histopathological feature to decrease the number of the attributes
(K et al., 2015)	Naive Bayes' (NB) and J48	the result showed that Naive Bayes produced less precision and true Positive rate as compared with J48 algorithm	The dataset contain just 230 instance , needs to be increased.
(Parikh et al., 2015)	artificial Neural Network and Support Vector Machine	the result which obtained from both classifiers ANN and SVM showed that the performance of ANN is more accurate for two hidden layers than SVM	To evaluate the performance of the classifiers, the dataset divided into 80-20%, across validation is recommended in such cases

Table (2.1) Analysis of the related works

2.5 Summary

This chapter presents four previous studies in diagnosing dermatology diseases .The first study use the Bayes net and Best First Search. The second study used Bayes net and Best First Search, the third study used Naive Bayes’ (NB) and J48. The fourth study used artificial Neural Network and Support Vector Machine.

Chapter Three

Methodology

3.1 Introduction

This chapter contains general information about the Weka software, classification and prediction, Data description and the algorithms that had used to classify the data.

3.2 Dataset Description

Two different dermatology datasets had been used in this research. One of them from the UCI website and the other is for Sudanese patients, it had been collected from Omdurman Military Hospital department of Dermatology. The Sudanese dataset had been built from the patient’s medical reports.

3.2.1 UCI Dermatology dataset description

This dataset collected by Bilkent University, department of Computer Engineering and Information Science. The number of Instances is 366 . It contains 34 attributes, 33 of which are nominal valued and one of them is linear.

There are six diseases in this dataset :

1- Psoriasis

2- seboreic dermatitis

3- lichen planus

4- pityriasis rosea

5- cronic dermatitis

6- pityriasis rubra pilaris.

Attribute Information:

Firstly : **Clinical Attributes:** (take values 0, 1, 2, 3)

1: erythema

2: scaling

3: definite borders

4: itching

5: koebner phenomenon

6: polygonal papules

7: follicular papules

8: oral mucosal involvement

9: knee and elbow involvement

10: scalp involvement

11: family history, (0 or 1)

12: Age (linear)

Secondly : **Histopathological Attributes**: (take values 0, 1, 2, 3)

1. melanin incontinence
2. eosinophils in the infiltrate
3. PNL infiltrate
4. fibrosis of the papillary dermis
5. exocytosis
6. acanthosis
7. hyperkeratosis
8. parakeratosis
9. clubbing of the rete ridges
10. elongation of the rete ridges
11. thinning of the suprapapillary epidermis
12. spongiform pustule
13. munro microabcess
14. focal hypergranulosis
15. disappearance of the granular layer
16. vacuolisation and damage of basal layer
17. spongiosis
18. saw-tooth appearance of retes
19. follicular horn plug
20. perifollicular parakeratosis
21. inflammatory monoluclear inflitrate
22. band-like infiltrate

Table 3. 1 Class Distribution for the UCI Dataset:

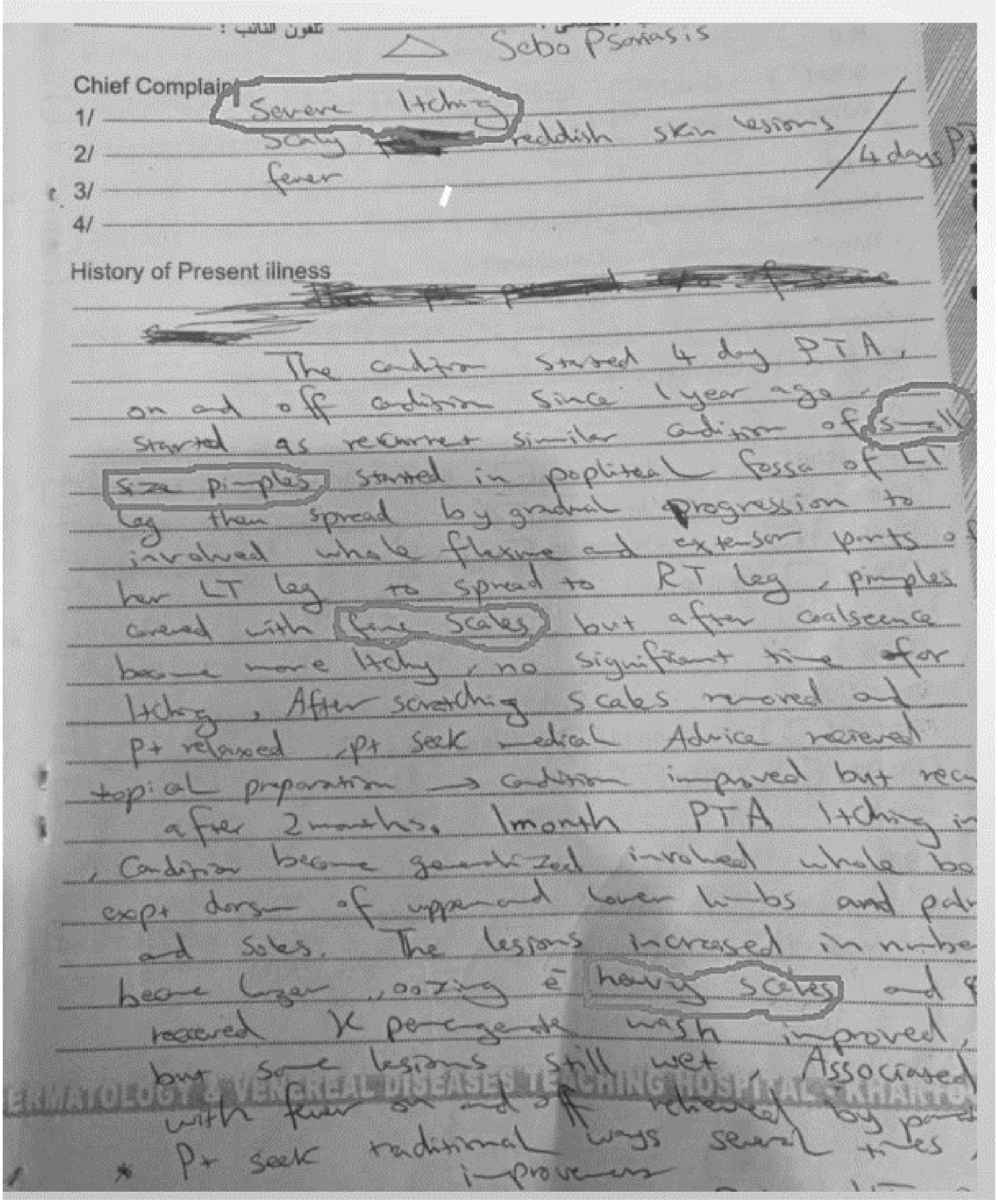
Class code	Class	Number of instances
1	Psoriasis	112
2	seboreic dermatitis	61
3	lichen planus	72
4	pityriasis rosea	49
5	cronic dermatitis	52
6	pityriasis rubra pilaris	20

3.2.2 Sudanese Dataset (SDD) Description

The data set had been built from the patients cards (see Appendix) in Omdurman Military Hospital department of Dermatology. The number of Instances is 1254 with 13 attributes. Twelve of them are nominal and one of them is linear. The patient's names were ignored. The family history feature has the value 1 if any of these diseases has been observed in the family, and 0 otherwise. The age feature simply represents the age of the patient. The clinical symptoms was given a degree in the range of 0 to 3. The zero indicates that the feature was not present, The Three indicates the largest amount possible, and 1, 2 indicate the relative intermediate values. In the following medical report example (Fig. 3.1) the sever itching symptom will be written 3 in the dataset .The small size pimple symptom will be written 1 in the dataset. The fine scales symptom will be written 1 .The heavy scales will written 3 in the dataset. The rest of the symptom which hadn't appear in the medical report will written 0.

The data set contains four diseases which are:

1- Psoriasis
2- seboreic dermatitis
3- lichen planus
4- cronic dermatitis



تلفون النائب :

△ Sebo Psoriasis

Chief Complaint

1/ Severe Itching

2/ Scaly [redacted] reddish skin lesions

3/ fever

4/

/4 days P

History of Present iliness

[redacted]

The condition started 4 day PTA, on and off condition since 1 year ago, started as recurrent similar condition of small size pimples started in popliteal fossa of LT leg then spread by gradual progression to involved whole flexure and extensor parts o her LT leg to spread to RT leg, pimples covered with fine scales but after coalsence become more Itchy, no significant time for Itching, After scratching scales removed and pt released, pt seek medical Advice received topical preparation → condition improved but recu after 2 months. 1month PTA Itching in, Condition become generalized involved whole bo expt dorsum of upper and lower limbs and pal and soles. The lesions increased in numbe become larger, oozing ē heavy scales and received K permanganate wash improved, but some lesions still wet, Associated with fever on and off relieved by para

* Pt seek traditional ways several times improvement

Fig. 3.1 example of medical report

Attribute Information:

Clinical Attributes:

1: erythema

2: scaling

3: definite borders

4: itching

5: koebner phenomenon

6: polygonal papules

7: follicular papules

8: oral mucosal involvement

9: knee and elbow involvement

10: scalp involvement

11: family history, (0 or 1)

12: Age (linear)

Table 3. 2 Class Distribution for the Sudanese Dataset:

Class	Number of instances
Psoriasis	368
seboreic dermatitis	290
lichen planus	298
cronic dermatitis	298

3.3 Data Preprocessing:

During the data creation, a lot of medical reports with missing values or old papers was ignored in order to get cleaned data free of defects. Such data will increase the data quality, raise the performance and get accurate results. Then the data had been converted from categorical in medical reports to nominal data (see Appendix A Fig. A.5 SDD)

Data mining algorithms may give poor results due to class imbalance problem, so the data already built with balance consideration in order to improve the accuracy, finally the data converted to csv format to be suitable for Weka software.

3.4 Classification and Prediction

Classification is the process of finding a model (or function) that describes and distinguishes data classes, in order to be able using this model to predict the class of objects whose class label is unknown

The new derived model depend on the analysis of a set of training data (data objects whose class label is known).The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks, there are many other methods for constructing classification models, such as naïve Bayesian classification, support vector machines, and k-nearest neighbor classification

3.4.1 Bayesian Classification

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given tuple belongs to a particular class. Bayesian classification is based on Bayes' theorem, Bayesian classifiers have exhibited high accuracy and speed when applied to large databases(Han and Kamber).

3.4.1.1 Bayes' Theorem

Bayes' theorem is named after Thomas Bayes, an English clergyman who did early work in probability and decision theory during the 18th century, the Theorem looking for the probability that tuple X belongs to class C, given that the attribute description of X are already known(Han and Kamber**).**

3.4.1.2 Naïve Bayesian classification

According to (Han and Kamber) Naive Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in a given data set. The algorithm uses Bayes theorem and assumes all attributes to be independent given the value of the class variable. Naïve Bayesian classifier is based on Bayes' theorem. The naïve Bayesian classifier, or simple Bayesian classifier, works as follows:

1) Let D be a training set of tuples and their associated class labels. As usual, each tuple is represented by an n-dimensional attribute vector, X = (x1,x2,...,xn), depicting n measurements made on the tuple from n attributes, respectively, A1,A2,...,An.
2) Suppose that there are m classes, C1,C2,...,Cm. Given a tuple, X, the classifier will Predict that X belongs to the class having the highest posterior probability, conditioned on X. That is, the naïve Bayesian classifier predicts that tuple X belongs to the class C_i if and only if

$$P(C_i|X) > P(C_j|X) \quad \text{for } 1 \le j \le m, j \ne i. \qquad 3.1$$

Thus we maximize $P(C_i |X)$.The class Ci for which $P(C_i |X)$ is maximized is

called the maximum posteriori hypothesis

$$P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)}. \qquad 3.2$$

3) As P(X) is constant for all classes, only $P(X| C_i)P(C_i)$ need be maximized. If the class prior probabilities are not known, then it is commonly assumed that the classes are equally likely, that is, P(C1) = P(C2) =···= P(Cm), and we would therefore

maximizeP(X|C_i).Otherwise,wemaximizeP(X|Ci)P(Ci).Notethattheclasspriorprobabilit iesmaybeestimatedbyP(C_i)=| C_i,D|/|D|,where|C_i,D|is the number of training tuples of class C_i in D.

4) Given data sets with many attributes, it would be extremely computationally expensive to compute P(X|Ci). In order to reduce computation in evaluating P(X|Ci), the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes). Thus,

$$\begin{aligned} P(X|C_i) &= \prod_{k=1}^{n} P(x_k|C_i) \\ &= P(x_1|C_i) \times P(x_2|C_i) \times \cdots \times P(x_n|C_i). \end{aligned} \quad 3.3$$

5) In order to predict the class label of X, P(X|Ci)P(Ci) is evaluated for each class Ci. The classifier predicts that the class label of tuple X is the class Ci if and only if

$$P(X|C_i)P(C_i) > P(X|C_j)P(C_j) \quad \text{for } 1 \leq j \leq m, j \neq i. \quad 3.4$$

3.4.2 Decision tree algorithm J48

According to (Patil and Sherekar, 2013) J48 classifier is a simple C4.5 decision tree for classification. It creates a binary tree. The decision tree approach is most useful in classification problem. Once the tree is built, it is applied to each tuple in the database and results in classification for that tuple.

While building a tree, J48 can predict the missing values based on what is known about the attribute values for the other records. The basic idea is to divide the data based on the attribute values for that item that are found in the training sample. J48 allows classification via either decision trees or rules generated from them.

3.4.3 Lazy Classifier

Lazy learners store the training instances and do no real work until classification time. Lazy learning is a learning method in which generalization beyond the training data is delayed until a query is made to the system where the system tries to generalize the training data before receiving queries.

The main advantage gained in employing a lazy learning method is that the target function will be approximated locally such as in the k-nearest neighbor algorithm.

Because the objective function is approximated locally for each query to the system, lazy learning systems can concurrently solve multiple problems and deal successfully with changes in the problem arena.

The disadvantages with lazy learning include the large space requirement to store the complete training dataset. Mostly noisy training data increases the case support unnecessarily, because no concept is made during the training phase and another disadvantage is that lazy learning methods are usually slower to evaluate, though this is joined with a faster training phase (Ms S. Vijayarani).

3.4.3.1 IBK (K - Nearest Neighbor):

As mention in (Ms S. Vijayarani)   IBK is a k-nearest-neighbor classifier that uses the same distance metric. The number of nearest neighbors can be specified explicitly in the object editor or determined automatically using leave-one-out cross-validation focus to an upper limit given by the specified value. A kind of different search algorithms can be used to speed up the task of finding the nearest neighbors. A linear search is the default but further options include KD-trees, ball trees, and so-called “cover trees”. The distance function used is a parameter of the search method. The remaining thing is the same as for IBL—that is, the Euclidean distance; other options include Chebyshev, Manhattan, and Minkowski distances. Predictions from more than one neighbor can be weighted according to their distance from the test instance and two different formulas are implemented for converting the distance into a weight.

The number of training instances kept by the classifier can be restricted by setting the window size option.

3.5 Weka Definition

Weka (WAIKATO ENVIRONMENT FOR KNOWLEDGE ANALYSIS) is a well-known machine learning software written in Java, developed at Waikato University in New Zealand. Application tools allows to identify hidden information from database. The Weka application contains a collection of visualization tools and algorithms for solving real-world data mining problems (Eshwari Girish Kulkarni 2016), WEKA accepts the data either in ARFF format or CSV format, Data types can be numeric, nominal, string and date. Numeric, string and date are case sensitive (Srivastava, 2014)

3.6 Summary

This chapter contains general information about the Weka software ,classification and prediction ,Data description and conversion from medical reports to the dataset .Also contain the algorithms that had used to classify the data .

Chapter Four

Results, Analysis and Discussion

4.1 Introduction:

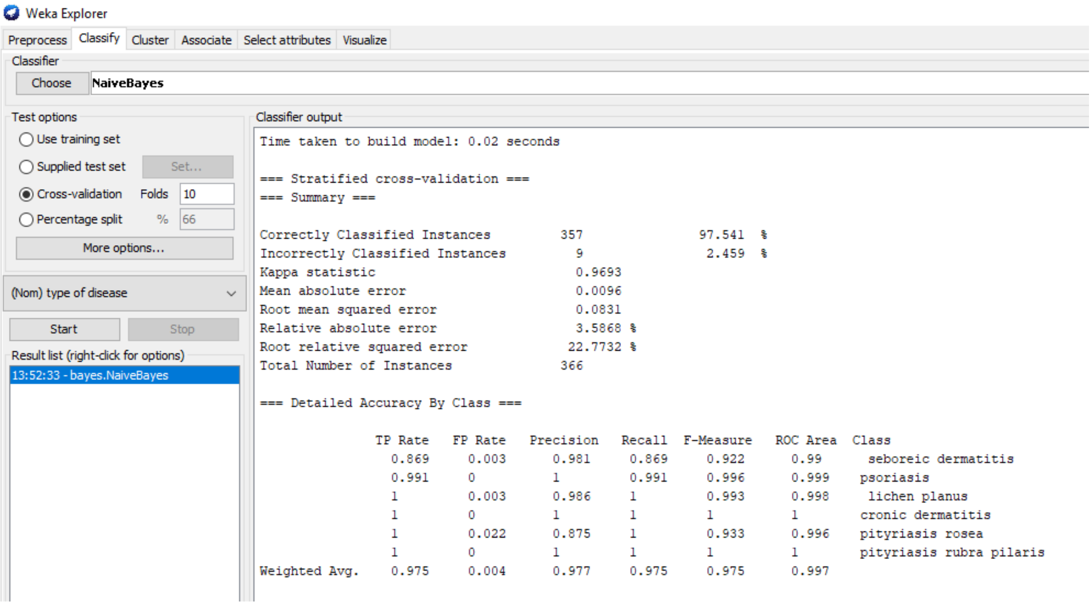
This chapter contain classification experiments, Results analysis and evaluation.

4.2 Experiments description

In the first experiments the Naïve Bayes, j48 and IBK algorithms were applied on the UCI dataset. Firstly using all the attributes (clinical and histopathological), secondly the histopathologic attributes was ignored because there are not available in the Sudanese dataset. In the second experiments three algorithms are applied on the Sudanese dermatology data set (SDD).

4.2.1 First Experiment:

In this experiment, the Naïve Bayes algorithm was applied on the UCI data set using all characteristics then only clinical characteristics were used to determine the effect of histological characteristics on the accuracy of the results. When we applied naïve Bayes on the UCI data set using all the features, the accuracy was 97.5%as shown in fig. 4.1



NaiveBayes

○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

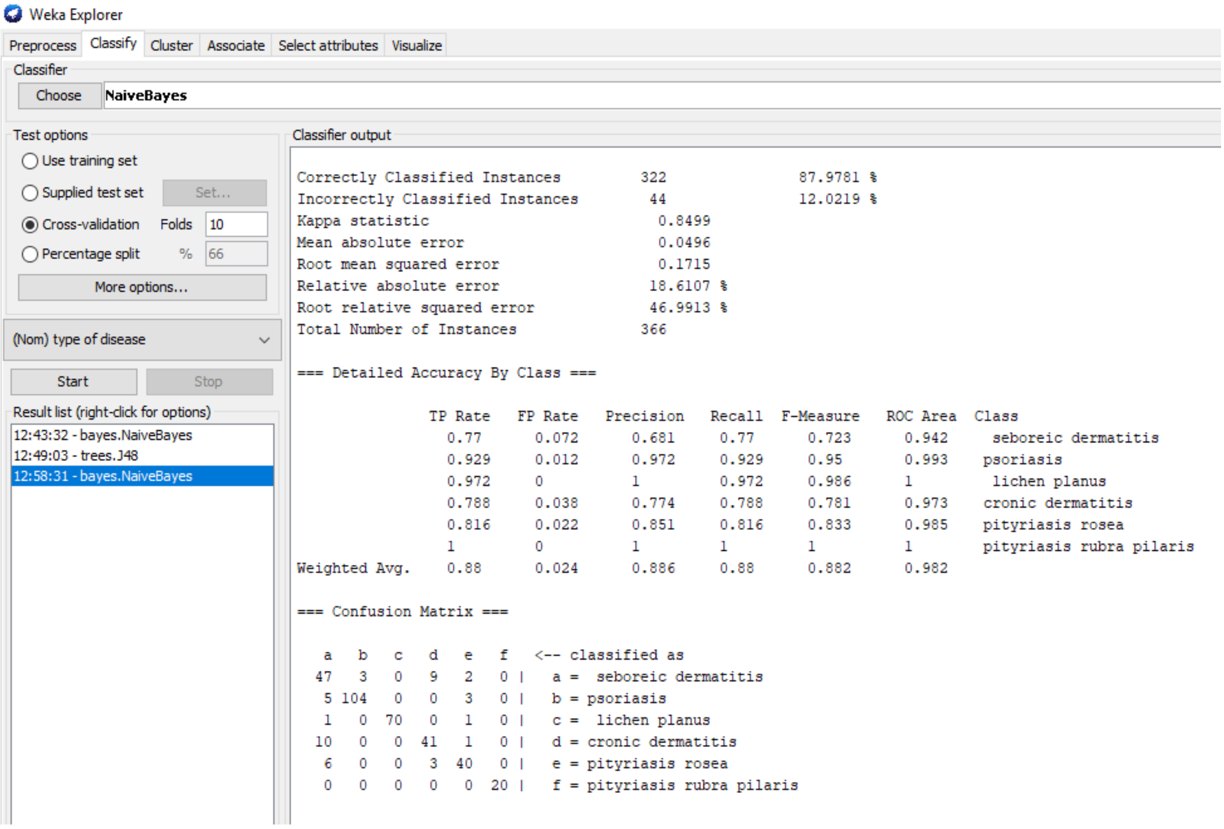
(Nom) type of disease

Result list (right-click for options)
13:52:33 - bayes.NaiveBayes

```
Time taken to build model: 0.02 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         357               97.541  %
Incorrectly Classified Instances         9                2.459  %
Kappa statistic                          0.9693
Mean absolute error                      0.0096
Root mean squared error                  0.0831
Relative absolute error                  3.5868 %
Root relative squared error             22.7732 %
Total Number of Instances              366

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.869     0.003     0.981      0.869     0.922      0.99      seboreic dermatitis
                 0.991     0         1          0.991     0.996      0.999    psoriasis
                 1         0.003     0.986      1         0.993      0.998      lichen planus
                 1         0         1          1         1          1        cronic dermatitis
                 1         0.022     0.875      1         0.933      0.996    pityriasis rosea
                 1         0         1          1         1          1        pityriasis rubra pilaris
Weighted Avg.    0.975     0.004     0.977      0.975     0.975      0.997
```

Fig. 4.1 Applying naïve byes classifier on UCI dataset (all attributes)

Then the naïve Bayes algorithm was applied on the UCI dataset, but this time the histopathological attributes was ignored. We just used the 12 clinical attributes, the obtained accuracy was 87.9 %. We observed that when we ignored the histopathological attributes the accuracy becomes lower than before (see Fig. 4.2).



Classifier: NaiveBayes

Test options
○ Use training set
○ Supplied test set
● Cross-validation Folds 10
○ Percentage split % 66

(Nom) type of disease

Result list (right-click for options)
12:43:32 - bayes.NaiveBayes
12:49:03 - trees.J48
12:58:31 - bayes.NaiveBayes

Classifier output

```
Correctly Classified Instances         322               87.9781 %
Incorrectly Classified Instances        44               12.0219 %
Kappa statistic                          0.8499
Mean absolute error                      0.0496
Root mean squared error                  0.1715
Relative absolute error                 18.6107 %
Root relative squared error             46.9913 %
Total Number of Instances              366

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
                 0.77     0.072    0.681     0.77     0.723      0.942    seboreic dermatitis
                 0.929    0.012    0.972     0.929    0.95       0.993   psoriasis
                 0.972    0        1         0.972    0.986      1        lichen planus
                 0.788    0.038    0.774     0.788    0.781      0.973   cronic dermatitis
                 0.816    0.022    0.851     0.816    0.833      0.985   pityriasis rosea
                 1        0        1         1        1          1       pityriasis rubra pilaris
Weighted Avg.    0.88     0.024    0.886     0.88     0.882      0.982

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
  47   3   0   9   2   0 |   a =  seboreic dermatitis
   5 104   0   0   3   0 |   b = psoriasis
   1   0  70   0   1   0 |   c =  lichen planus
  10   0   0  41   1   0 |   d = cronic dermatitis
   6   0   0   3  40   0 |   e = pityriasis rosea
   0   0   0   0   0  20 |   f = pityriasis rubra pilaris
```

4.2 Applying naïve byes classifier on UCI dataset (Only Clinical attributes)

Then we apply the j48 and IBK on the UCI dataset firstly with all the attributes, secondly with clinical attributes, the accuracy was as in table 4.1 and chart 4.1

Table 4. 1Accuracy of the three algorithms when applied on UCI dataset

Attributes / Algorithm	Accuracy %	
	clinical + Histopathological	**Clinical**
Naïve Bayes	97.541 %	87.9781 %
J48	95.9016 %	86.8852 %
IBK	95.3552 %	81.9672 %

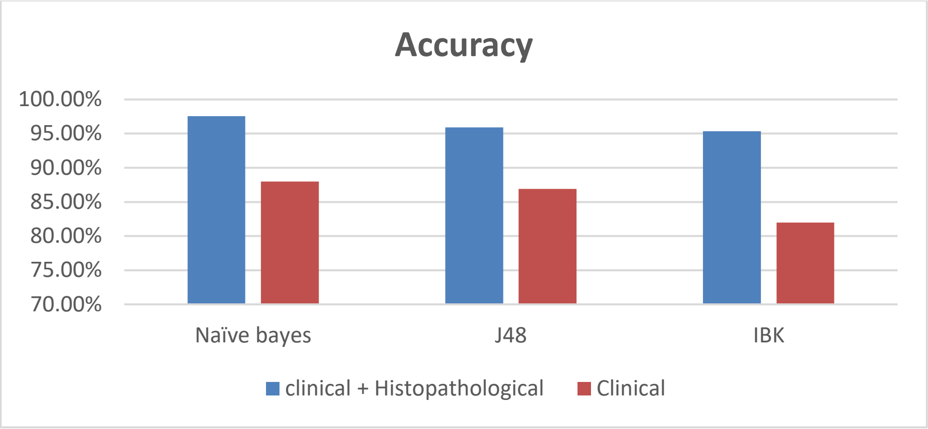


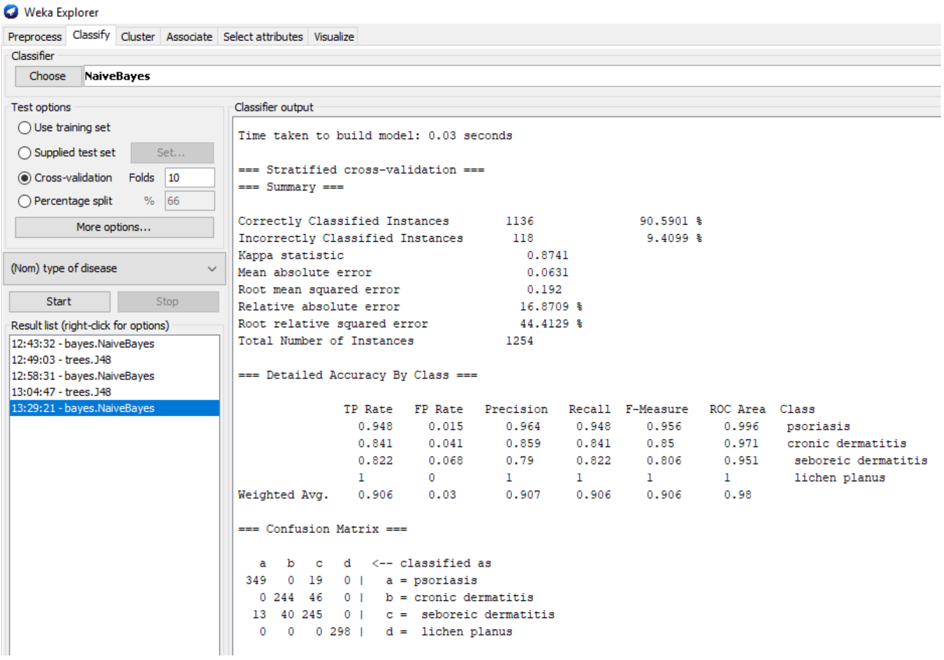
Figure 4. 1Accuracy of the three algorithms when applied on UCI dataset

4.2.2 Second Experiments

The main difference between the UCI data set and Sudanese data set is in the histopathological attributes where we did not find them written in the Sudanese patients' medical report so we could not include them in the data. Therefore, the Sudanese data contains only the clinical attributes.

In this experiments we applied the three algorithms on the Sudanese patient’s data in spite of lack of histopathological attributes to observe the effect of its absence on the accuracy of the model.

Firstly a model was built using Naïve Bayes, the accuracy was 90.5 % (see Fig. 4.3). As observed from the confusion Matrix(Table 4.2) that the model did not classify 46 instance of cronic dermatitis correctly. But classified them as seboriec Dermmatatis and vice versa there are 40 Instance of seboreic have been classified as cronic dermatities , This indicates the high similarity of the symptoms of this two diseases.



Choose NaiveBayes

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) type of disease

12:43:32 - bayes.NaiveBayes
12:49:03 - trees.J48
12:58:31 - bayes.NaiveBayes
13:04:47 - trees.J48
13:29:21 - bayes.NaiveBayes

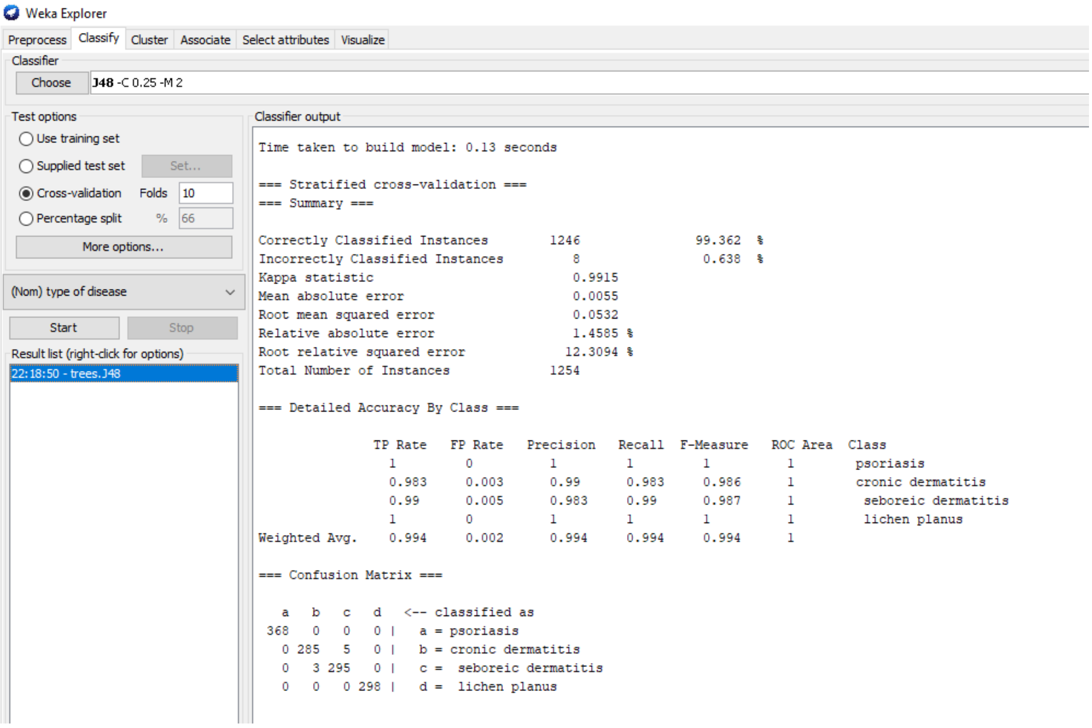
```
Time taken to build model: 0.03 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1136               90.5901 %
Incorrectly Classified Instances       118                9.4099 %
Kappa statistic                          0.8741
Mean absolute error                      0.0631
Root mean squared error                  0.192
Relative absolute error                 16.8709 %
Root relative squared error             44.4129 %
Total Number of Instances             1254

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.948     0.015     0.964     0.948     0.956      0.996    psoriasis
                 0.841     0.041     0.859     0.841     0.85       0.971    cronic dermatitis
                 0.822     0.068     0.79      0.822     0.806      0.951     seboreic dermatitis
                 1         0         1         1         1          1         lichen planus
Weighted Avg.    0.906     0.03      0.907     0.906     0.906      0.98

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 349   0  19   0 |   a = psoriasis
   0 244  46   0 |   b = cronic dermatitis
  13  40 245   0 |   c =  seboreic dermatitis
   0   0   0 298 |   d =  lichen planus
```

Fig. 4.3 Applying Naïve bayes on Sudanese dataset

Table 4. 2 Confusion matrix when apply naïve byes on Sudanese dataset

A	B	C	D	classified as
349	0	19	0	A = Psoriasis
0	244	46	0	B = cronic dermatitis
13	40	245	0	C = seboreic dermatitis
0	0	0	298	D = lichen planus

Secondly The **J48** algorithm was applied on the Sudanese dataset, the accuracy was 99.36% (see Fig. 4.4)



J48 -C 0.25 -M 2

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) type of disease

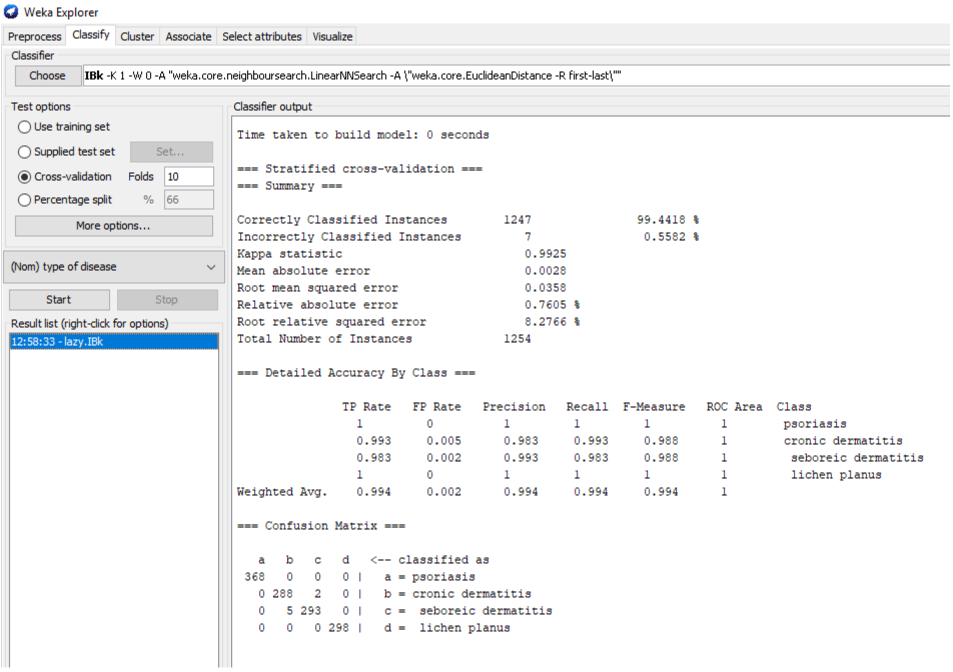
22:18:50 - trees.J48

```
Time taken to build model: 0.13 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1246               99.362  %
Incorrectly Classified Instances         8                0.638  %
Kappa statistic                          0.9915
Mean absolute error                      0.0055
Root mean squared error                  0.0532
Relative absolute error                  1.4585 %
Root relative squared error             12.3094 %
Total Number of Instances             1254

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1          1        1          1        psoriasis
                 0.983     0.003      0.99       0.983    0.986      1        cronic dermatitis
                 0.99      0.005      0.983      0.99     0.987      1         seboreic dermatitis
                 1         0          1          1        1          1         lichen planus
Weighted Avg.    0.994     0.002      0.994      0.994    0.994      1

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 368   0   0   0 |   a = psoriasis
   0 285   5   0 |   b = cronic dermatitis
   0   3 295   0 |   c =  seboreic dermatitis
   0   0   0 298 |   d =  lichen planus
```

Fig. 4.4 Applying J48 on the Sudanese dataset

In this experiment the accuracy was better than before and there was improvement in the confusion matrix, only five diseases from cronic dermatitis not classified correctly and 3 instances from seboriec dermatitis classified as cronic dermatitis.

Finally the **Lazy** IBK algorithm applied on the Sudanese dataset (SDD), the accuracy was 99.44%(see Fig. 4.5).



IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""

● Cross-validation Folds 10

○ Percentage split % 66

(Nom) type of disease

12:58:33 - lazy.IBk

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        1247               99.4418 %
Incorrectly Classified Instances         7                0.5582 %
Kappa statistic                          0.9925
Mean absolute error                      0.0028
Root mean squared error                  0.0358
Relative absolute error                  0.7605 %
Root relative squared error              8.2766 %
Total Number of Instances             1254

=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 1         0          1          1        1          1        psoriasis
                 0.993     0.005      0.983      0.993    0.988      1        cronic dermatitis
                 0.983     0.002      0.993      0.983    0.988      1         seboreic dermatitis
                 1         0          1          1        1          1         lichen planus
Weighted Avg.    0.994     0.002      0.994      0.994    0.994      1

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 368   0   0   0 |   a = psoriasis
   0 288   2   0 |   b = cronic dermatitis
   0   5 293   0 |   c =  seboreic dermatitis
   0   0   0 298 |   d =  lichen planus
```

Fig 4.5 Applying IBK algorithm on the Sudanese dataset

In this experiment the accuracy was better than before and the improvement in the confusion matrix was very satisfactory, only five diseases from seboriec dermatitis not classified correctly and just two instances from cronic dermatitis classified as seboriec dermatitis.

Table 4. 3confusion matrix when apply IBK on Sudanese dataset

A	B	c	D	classified as	
368	0	0	0	a =	Psoriasis
0	288	2	0	b =	cronic dermatitis
0	5	293	0	c =	seboreic dermatitis
0	0	0	298	d =	lichen planus

In the recent experiment, we noticed that lazy IBK was the best algorithm where we get the highest accuracy, less time and best confusion matrix.

The following chart explain the accuracy of the three models

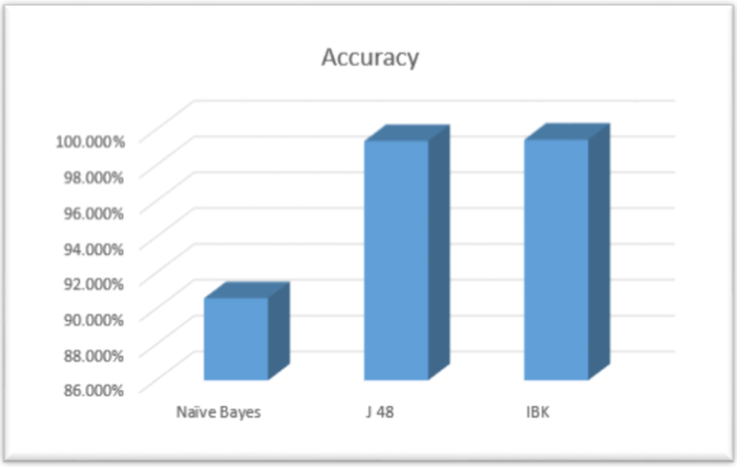


Figure 4.6 Compare the accuracy of the three models (second experiment)

The following table show the accuracy measures for the three classification techniques, and the time which taken to build the models

Table 4. 4Accuracy measures and time for the three models (second experiment)

Algorithm	Correctly Classified Instances (%) value	Incorrectly Classified Instances (%) value	Accuracy	Time taken to build the model
Naïve bays	1136	118	90.6%	0.05 seconds
J 48	1246	8	99.36%	0.29 seconds
IBK	1247	7	99.44%	0 seconds

4.3 Summary

This chapter contain two experiments.in the first experiments the Naïve Bayes, j48 and IBK algorithms were applied on the UCI dataset. Firstly using all the attributes (clinical and histopathological), secondly use just the clinical attributes .In this experiments all the results show high accuracy when used all attributers. When the histopathologic attributes was ignored all the results was lower than before.

In the second experiments the three algorithms were applied on the SDD using just the clinical attributes. The results showed that lazy IBK was the best algorithm .It gives the highest accuracy, less time and best confusion matrix.

Chapter Five

Conclusion and Recommendation

5.1 Introduction:

This chapter explains the research conclusion and some recommendation for futures consideration.

5.2 Conclusion:

The purpose of this research is to support doctors in diagnosing four dermatology diseases which have high similarity in clinical symptoms. Also the research aimed to use real Sudanese data which had been collected from scratch, and apply three data mining algorithms naive Bayes, J48 and IBK.

Medical data set must be correct and very accurate, the major challenges in this research was the data collection stage because of the difficulty of reading the doctors' hand writing and extracting data from medical reports , there for this data gathering cannot be accurate unless the some doctors assist us in filling out all this data set .

Two data sets had been used in this research, one from UCI repository website and the other is a real data collected from Omdurman military hospital, the attributes in these two data sets were different, the UCI data set have two kinds of attributes clinical attributes and Histopathological attributes but for Sudanese data set only clinical attributes was available , the Histopathological attributes didn't included in the medical reports.

After applying the three algorithms on the two datasets, the conclusion explain that the IBK (K NN) did not produce good results for the UCI dataset where a relatively large number of attributes are present, so the IBK algorithm did not produce good results in the case of many attributes and few data, but when the same algorithm applied on the Sudanese data the best results was gained where the instance number are more than the UCI dataset Instance, furthermore the attributes were fewer.

5.3 Recommendation

This was the first attempt to build dermatology data set from Omdurman military hospital. Therefore, we recommend increasing this data from Khartoum Hospital and integrate them into a large Sudanese dataset for future studies.

- Building real Sudanese medical data sets for various other diseases
- Apply data mining in medical domain as general, to assist doctors in making decisions,
- applying these algorithms on other skin diseases that have similar symptoms, such as eczema.
- We recommend also adding some features to the data set such as patient occupation, residency and gender, which were available in the Sudanese medical reports, maybe these features have an impact on the type of diagnosis.

References

ESHWARI GIRISH KULKARNI , R. B. K., PHD 2016. WEKA Powerful Tool in Data Mining. *International Journal of Computer Applications (0975 – 8887)*, 6.

HAN, J. & KAMBER, M. 2006. Data Mining Concepts and Techniques *second edition*.

K, M. K., K, S. & P, S. 2015 Data Mining in Dermatological Diagnosis: A Method for Severity Prediction *International Journal of Computer Applications (0975 – 8887)* 117 11-14.

K., M., K. SANKARANARAYANAN & SEENAP 2014 Prediction of Different Dermatological Conditions Using Naïve Bayesian Classification *International Journal of Advanced Research in Computer Science and Software Engineering* 4**,** 864 - 868

MADHURARAMBHAJANI1, DEEPANKER2, W. & PATHAK3, N. 2015. Classification of Dermatology Diseases through Bayes net and Best First Search *International Journal of Advanced Research in Computer and Communication Engineering* 4.

MS S. VIJAYARANI , M. M. M. Comparative Analysis of Bayes and Lazy classification algorithms.pdf>. *International Journal of Advanced Research in Computer and Communication Engineering,* Vol. 2 7.

PARIKH, K. S., SHAH, T. P., KOTA, R. & VORA, R. 2015. Diagnosing Common Skin Diseases using Soft Computing Techniques. *International Journal of Bio-Science and Bio-Technology,* 7**,** 275-286.

PATIL, T. R. & SHEREKAR, M. S. S. 2013. Performance Analysis of Naive Bayes and J48 Classification Algorithm for Data Classification *International Journal Of Computer Science And Applications* 6.

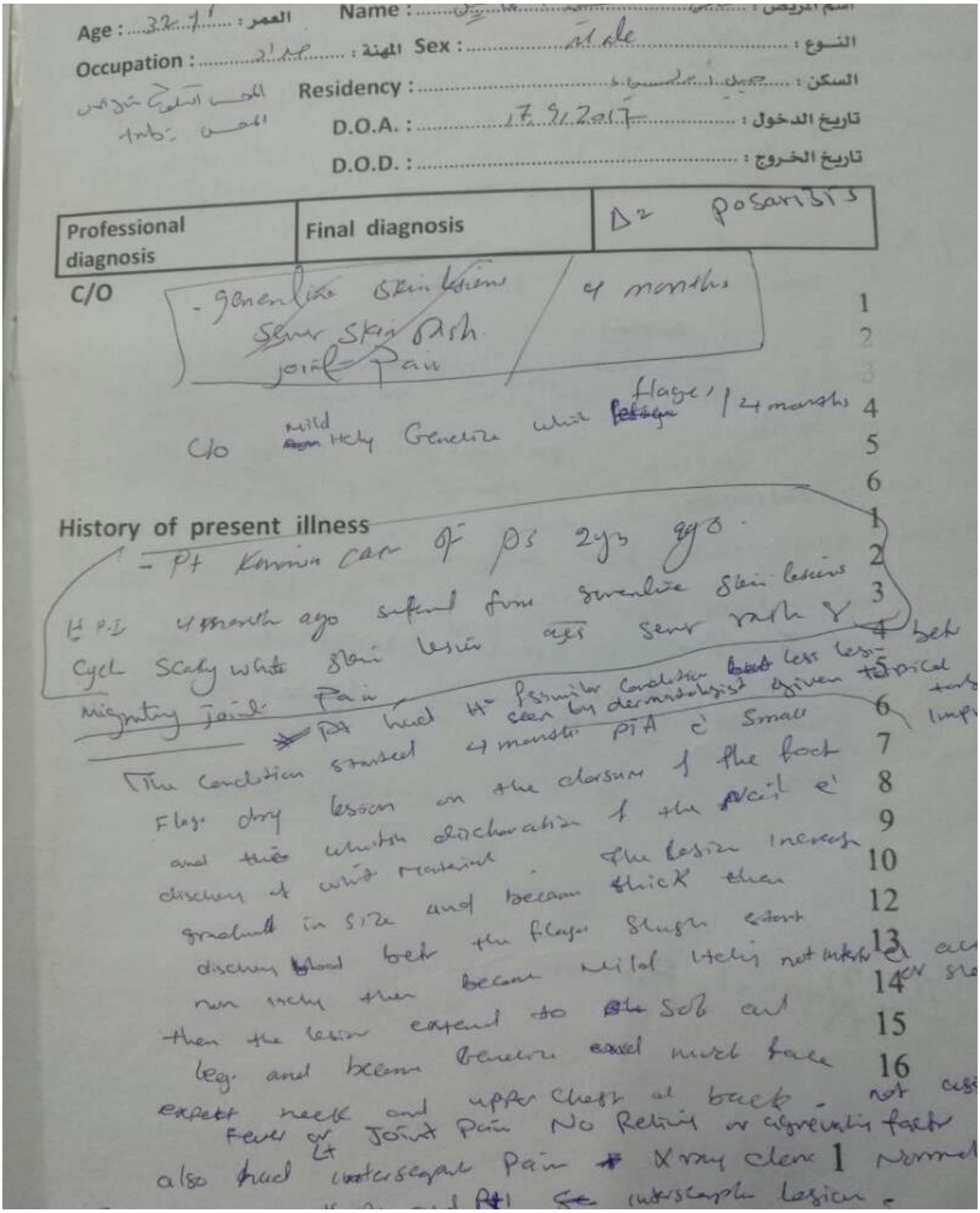
SISTLA KARTHIK*, B. S. R. M. V., TARUN TEJ K**** 2017 PREDICTION OF DERMATOLOGICAL CONDITION USING NAÏVE BAYESIAN CLASSIFICATION *International Journal of Pharmacy & Technology*
SRIVASTAVA, S. 2014. Weka: A Tool for Data preprocessing, Classification,Ensemble, Clustering and Association Rule Mining. *International Journal of Computer Applications (0975 – 8887),* 88.

THORAT1, S. & KUTE2, S. 2014. Medical Data Mining Life Cycle and its Role in Medical Domain. *Surabhi Thorat et al, / (IJCSIT) International Journal of Computer Science and Information Technologies,* Vol. 5 (4).

WASAN1, S. K., BHATNAGAR2, V., KAUR1*, H. & 2006 THE IMPACT OF DATA MINING TECHNIQUES ON MEDICAL DIAGNOSTICS *Data Science Journal,* 5.

Appendix A

Some examples for Sudanese real medical reports



Age : 32 y العمر : Name : [illegible] اسم المريض :

Occupation : [illegible] المهنة : Sex : Male النــوع :

[illegible] Residency : [illegible] السكن :

D.O.A. : 17.9.2017 تاريخ الدخول :

D.O.D. : تاريخ الخـروج :

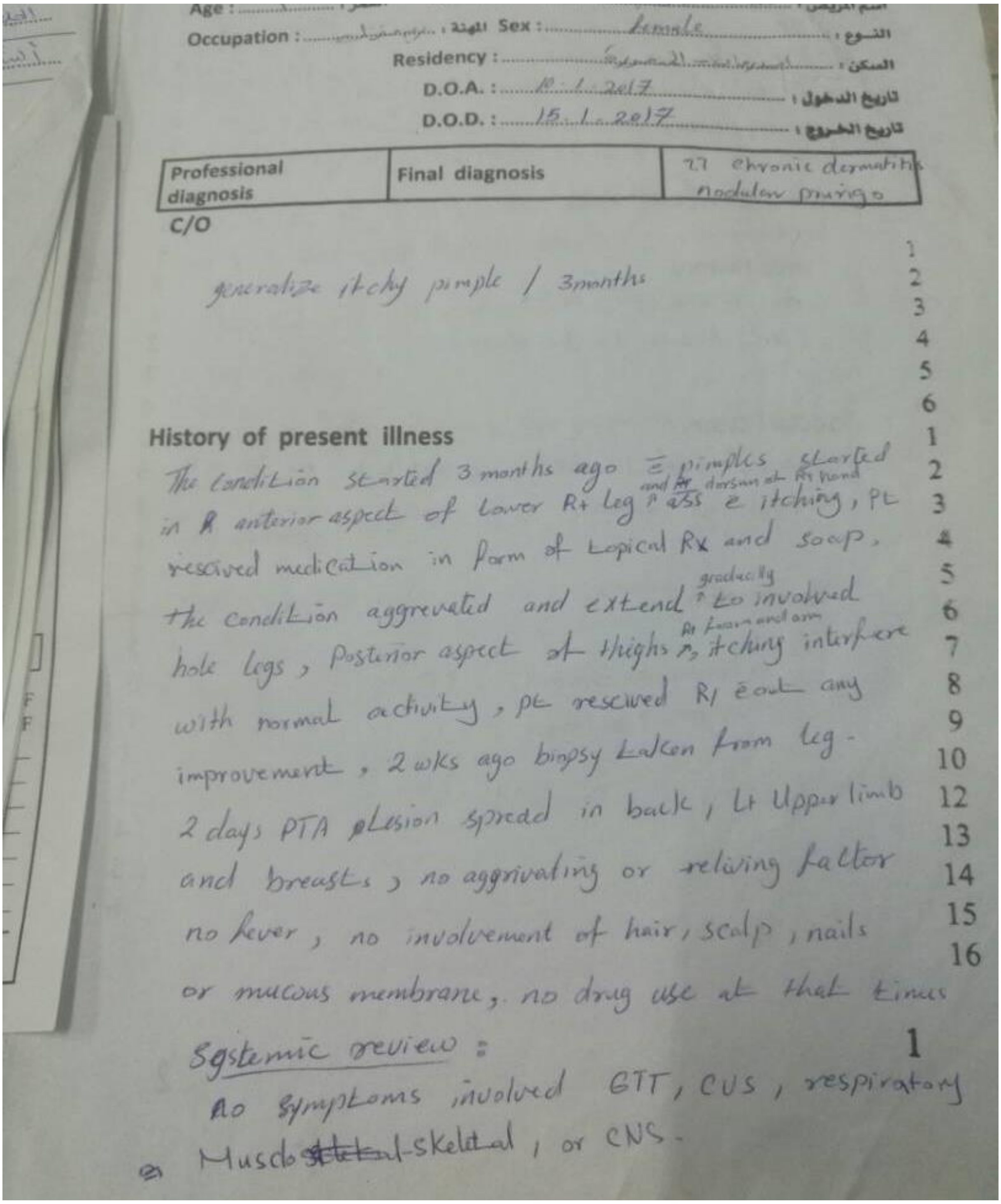
Professional diagnosis	Final diagnosis	Δ² posarisis

C/O - generlize skin lesions / 4 months
~~sever~~ skin rash
joint pain

C/O ~~non~~ mild itchy Generlize white ~~lesion~~ flage, | 4 months

History of present illness
- Pt known case of ps 2ys ago.
H P.I 4 month ago sufferd from generlize skin lesions ~~ages~~ sever rash & bet
cycle scaly white skin lesion ~~but~~ less lesi- topical
migraty joint pain
* Pt had H° of similar condition seen by dermatologist given topical [illegible] improve
The condition started 4 months PTA c̄ small flage dry lesson on the dorsum of the foot
and the whitsh dischoration of the nail c̄ discharge of white material. The lesion increase
gradually in size and became thick then the flage slught start discharge ~~blood~~ bet
became mild itchy not [illegible] c̄ [illegible] non itchy then [illegible] the lesion extend to ~~the~~ sole and
leg. and became generlize ~~and~~ invole face
expect neck and upper chest al back - not [illegible]
fever or joint pain No Relieving or aggravating factor
also had Lt interscapular pain * X ray chest Normal
[illegible] Rt [illegible] interscapular lesion.

Fig. A.1 Medical report



Age : .... العمر : .... اسم المريض :

Occupation : [illegible] المهنة : Sex : female النوع :

Residency : [illegible] السكن :

D.O.A. : 10. 1. 2017 تاريخ الدخول :

D.O.D. : 15. 1. 2017 تاريخ الخروج :

Professional diagnosis	Final diagnosis	?? chronic dermatitis nodular prurigo

C/O

generalize itchy pimple / 3months

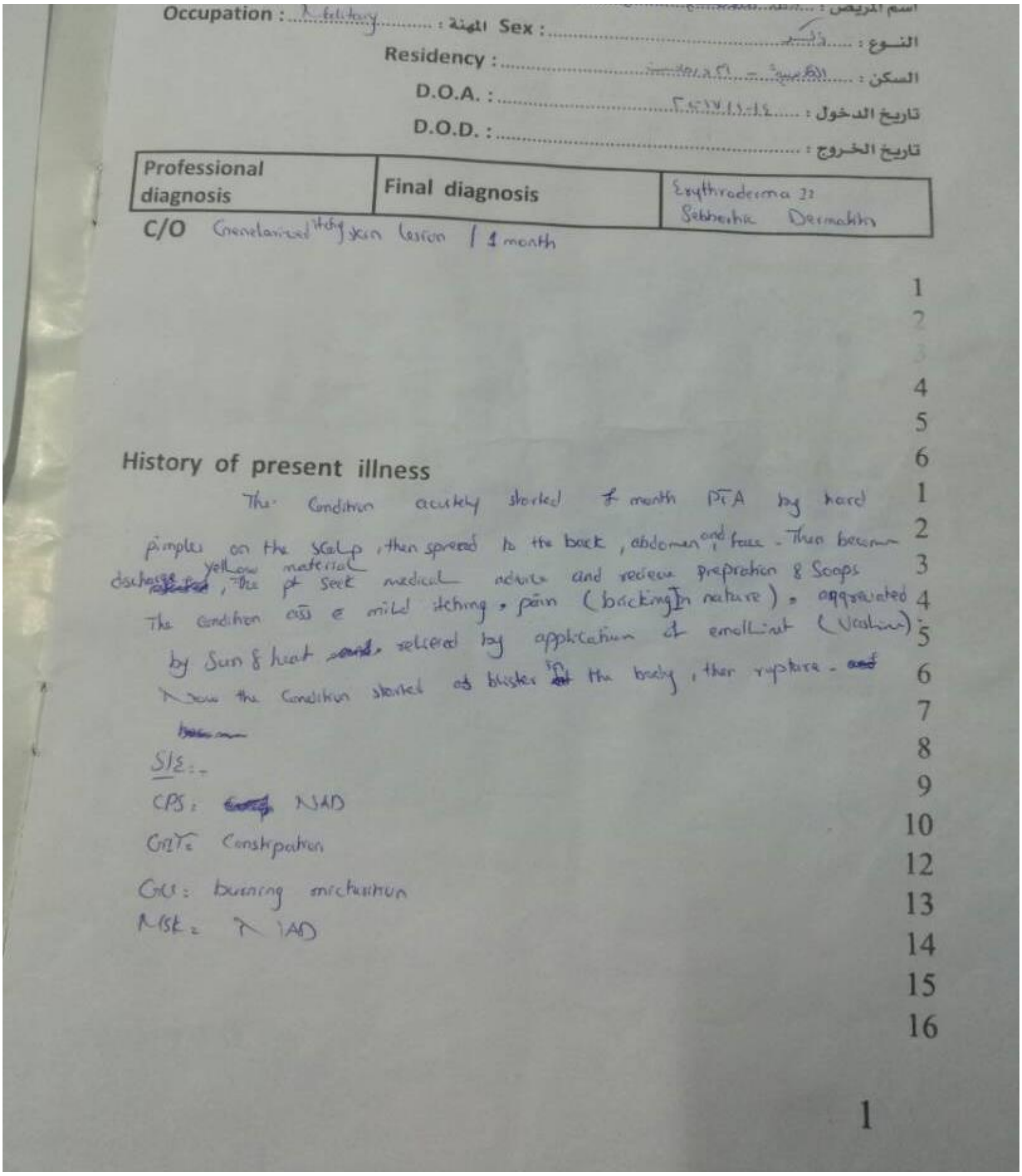
History of present illness

The condition started 3 months ago ē pimples started in R anterior aspect of lower Rt leg and Rt dorsum of Rt hand p ass ē itching, Pt rescived medication in form of topical Rx and soap, the condition aggrevated and extend gradually to involved hole legs, Posterior aspect of thighs Rt forearm and arm, itching interfere with normal activity, pt rescived R/ ēout any improvement, 2 wks ago biopsy taken from leg. 2 days PTA lesion spread in back, Lt Upper limb and breasts, no aggrivating or reliving factor no fever, no involvement of hair, scalp, nails or mucous membrane, no drug use at that times

Sgstemic review :

no symptoms involved GIT, CVS, respiratory Muscl~~stetal~~-skeletal, or CNS.

Fig. A.2 Medical report



اسم المريض : ...

Occupation : Military : المهنة Sex : ذكر : النـوع

Residency : [illegible] : السكن

D.O.A. : ٢٠١٧.١١.١٤ : تاريخ الدخول

D.O.D. : ...... : تاريخ الخروج

Professional diagnosis	Final diagnosis	Erythroderma 2ry Sebborhic Dermatitis

C/O Generalized itchy skin lesion / 1 month

History of present illness

The Condition acutely started 1 month PTA by hard pimples on the scalp, then spread to the back, abdomen and face. Then become discharge ~~[illegible]~~ yellow material, the pt seek medical advice and recieve prepration & Soaps The condition ass c mild itching, pain (bricking In nature), aggravated by Sun & heat ~~and~~, relieved by application of emollient (Vasline). Now the Condition started of blister in ~~at~~ the body, then rupture. ~~and~~ ~~becom~~

S/E:-

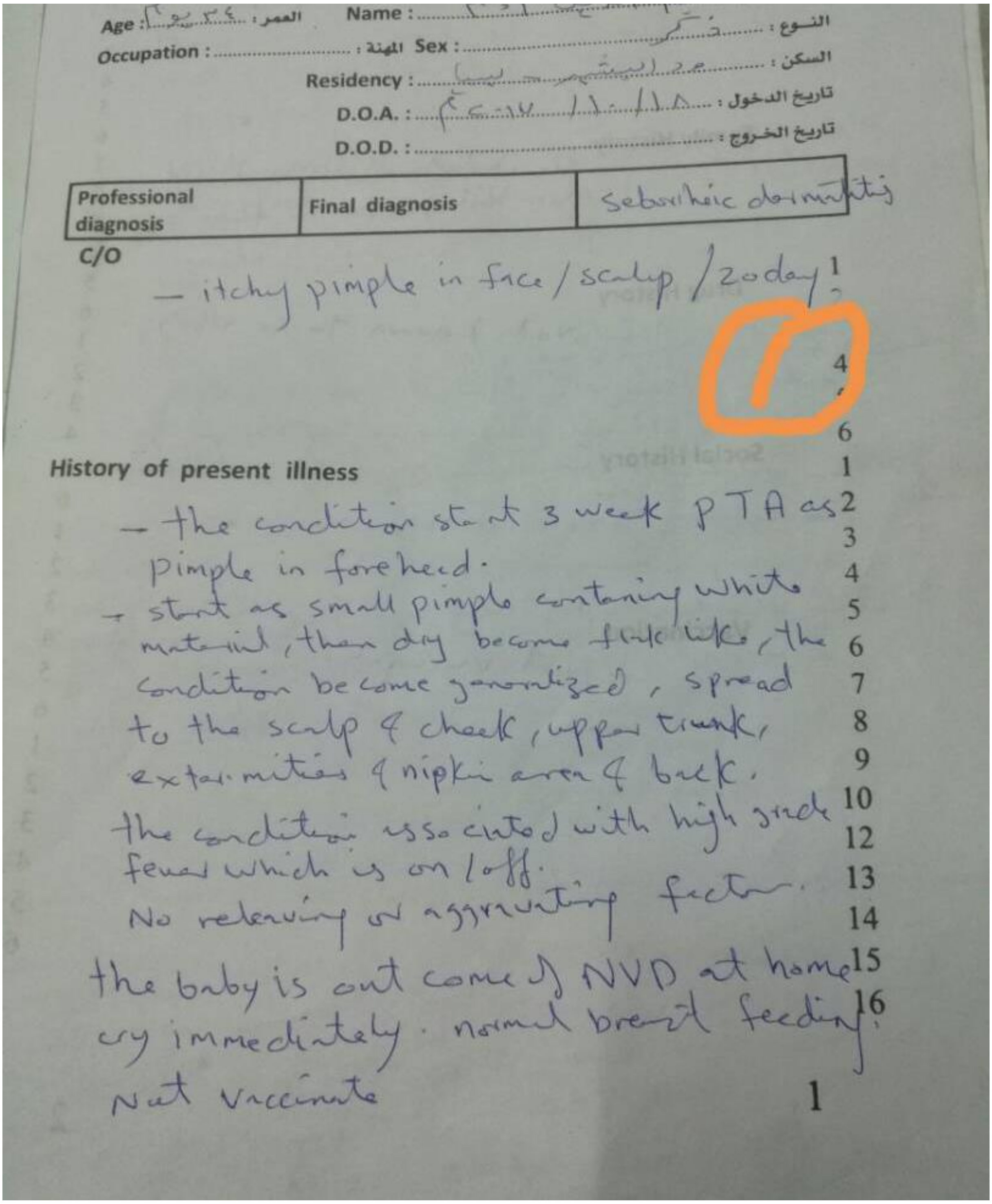
CPS: ~~[illegible]~~ NAD

GIT: Constipation

GU: burning micturition

MSK: N IAD

Fig. A.3 Medical report



Age : ٣٤ يوم العمر : Name : النـوع : ذكر

Occupation : المهنة : Sex : السكن : [illegible] ليبيا

Residency :

D.O.A. : ١٨/١٠/٢٠١٧م تاريخ الدخول :

D.O.D. : تاريخ الخروج :

Professional diagnosis	Final diagnosis	Seborrheic dermatitis

C/O

- itchy pimple in face / scalp / 20 day.

History of present illness

- the condition start 3 week PTA as pimple in forehead.
- start as small pimple containing white material, then dry become flake like, the condition become generalized, spread to the scalp & cheek, upper trunk, extarmities & nipkin area & back.

the condition associated with high grade fever which is on/off.

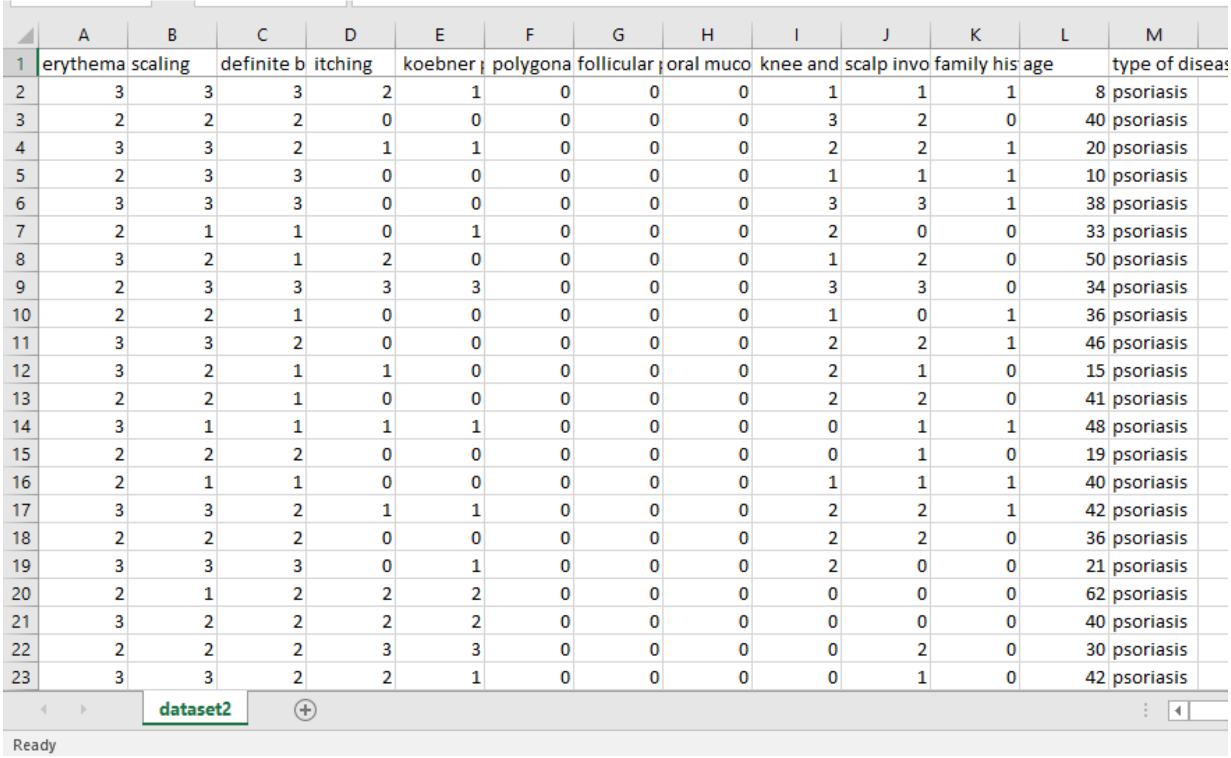
No releaving or aggravating factor.

the baby is out come of NVD at home cry immediately. normal breast feeding.

Not vaccinate

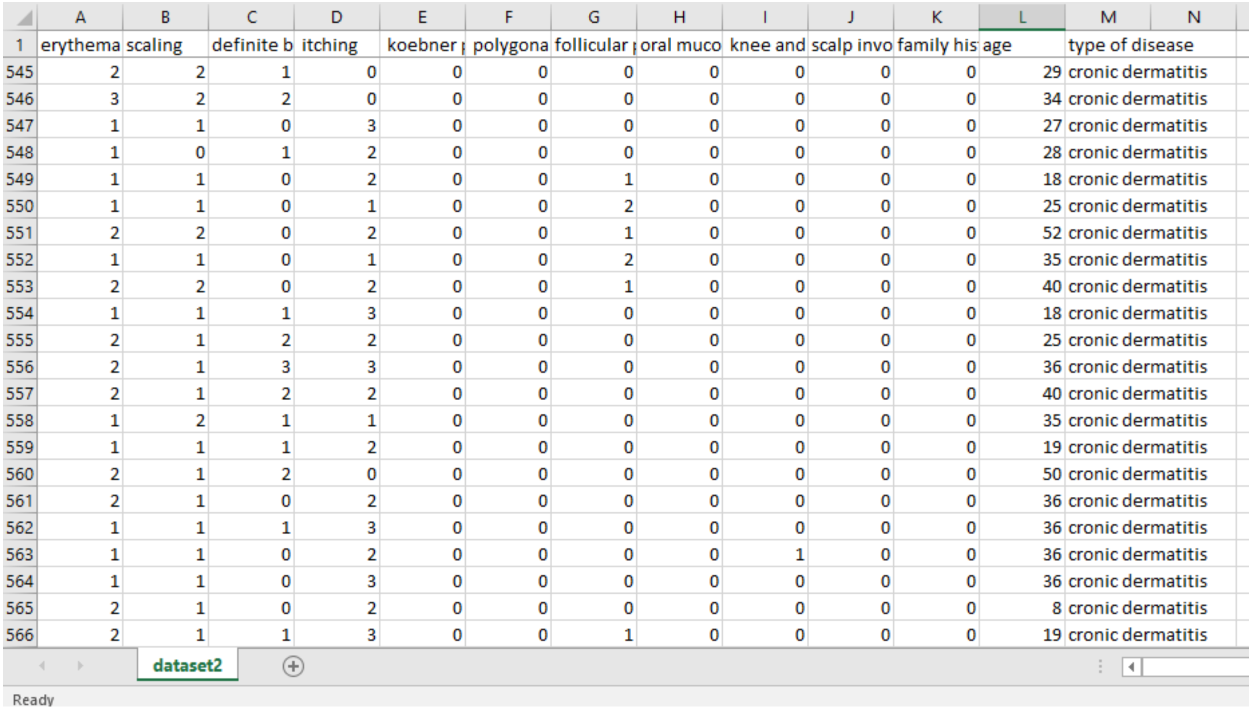
Fig. A.4 Medical report

Sample of Sudanese Dermatology Dataset (SDD)



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	erythema	scaling	definite b	itching	koebner	polygona	follicular	oral muco	knee and	scalp invo	family his	age	type of diseas
2	3	3	3	2	1	0	0	0	1	1	1	8	psoriasis
3	2	2	2	0	0	0	0	0	3	2	0	40	psoriasis
4	3	3	2	1	1	0	0	0	2	2	1	20	psoriasis
5	2	3	3	0	0	0	0	0	1	1	1	10	psoriasis
6	3	3	3	0	0	0	0	0	3	3	1	38	psoriasis
7	2	1	1	0	1	0	0	0	2	0	0	33	psoriasis
8	3	2	1	2	0	0	0	0	1	2	0	50	psoriasis
9	2	3	3	3	3	0	0	0	3	3	0	34	psoriasis
10	2	2	1	0	0	0	0	0	1	0	1	36	psoriasis
11	3	3	2	0	0	0	0	0	2	2	1	46	psoriasis
12	3	2	1	1	0	0	0	0	2	1	0	15	psoriasis
13	2	2	1	0	0	0	0	0	2	2	0	41	psoriasis
14	3	1	1	1	1	0	0	0	0	1	1	48	psoriasis
15	2	2	2	0	0	0	0	0	0	1	0	19	psoriasis
16	2	1	1	0	0	0	0	0	1	1	1	40	psoriasis
17	3	3	2	1	1	0	0	0	2	2	1	42	psoriasis
18	2	2	2	0	0	0	0	0	2	2	0	36	psoriasis
19	3	3	3	0	1	0	0	0	2	0	0	21	psoriasis
20	2	1	2	2	2	0	0	0	0	0	0	62	psoriasis
21	3	2	2	2	2	0	0	0	0	0	0	40	psoriasis
22	2	2	2	3	3	0	0	0	0	2	0	30	psoriasis
23	3	3	2	2	1	0	0	0	0	1	0	42	psoriasis

Fig. A.5 SDD



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	erythema	scaling	definite b	itching	koebner	polygona	follicular	oral muco	knee and	scalp invo	family his	age	type of disease	
545	2	2	1	0	0	0	0	0	0	0	0	29	cronic dermatitis	
546	3	2	2	0	0	0	0	0	0	0	0	34	cronic dermatitis	
547	1	1	0	3	0	0	0	0	0	0	0	27	cronic dermatitis	
548	1	0	1	2	0	0	0	0	0	0	0	28	cronic dermatitis	
549	1	1	0	2	0	0	1	0	0	0	0	18	cronic dermatitis	
550	1	1	0	1	0	0	2	0	0	0	0	25	cronic dermatitis	
551	2	2	0	2	0	0	1	0	0	0	0	52	cronic dermatitis	
552	1	1	0	1	0	0	2	0	0	0	0	35	cronic dermatitis	
553	2	2	0	2	0	0	1	0	0	0	0	40	cronic dermatitis	
554	1	1	1	3	0	0	0	0	0	0	0	18	cronic dermatitis	
555	2	1	2	2	0	0	0	0	0	0	0	25	cronic dermatitis	
556	2	1	3	3	0	0	0	0	0	0	0	36	cronic dermatitis	
557	2	1	2	2	0	0	0	0	0	0	0	40	cronic dermatitis	
558	1	2	1	1	0	0	0	0	0	0	0	35	cronic dermatitis	
559	1	1	1	2	0	0	0	0	0	0	0	19	cronic dermatitis	
560	2	1	2	0	0	0	0	0	0	0	0	50	cronic dermatitis	
561	2	1	0	2	0	0	0	0	0	0	0	36	cronic dermatitis	
562	1	1	1	3	0	0	0	0	0	0	0	36	cronic dermatitis	
563	1	1	0	2	0	0	0	0	1	0	0	36	cronic dermatitis	
564	1	1	0	3	0	0	0	0	0	0	0	36	cronic dermatitis	
565	2	1	0	2	0	0	0	0	0	0	0	8	cronic dermatitis	
566	2	1	1	3	0	0	1	0	0	0	0	19	cronic dermatitis	

Fig. A.6 SDD