Periodicals of Engineering and Natural Sciences ISSN 2303-4521
Vol. 7, No. 4, December 2019, pp.1801-1811

Hair analysis based on medical history and spatial-temporal data

Ahmed Mahdi Abdulkadium [1], Raid Abd Alreda Shekan [2], Ali Abdulbaqi Abdulazeez [3]
[1] Computer Center/Al-Qasim Green Universit
[2] College of education/University of Babylon
[3] Education Ministry/General of Basrah Education

ABSTRACT

Over the course of time, machine learning has improved the data analysis technique such as face detection and recognition. Many machine learning researches have been implemented in medical treatments. This concept which is proposed is inspired from different aspects of hair scalp and other factors. Spatial-temporal data is very useful in weather forecasting and satellite image analysis. This technique is implemented to capture necessary data from hair follicle images. Hair is also a subject of human body. There are many factors which can be used to determine health of hair. All these factors including spatial-temporal images, gender, age and hair style are used to predict health of hair. This paper presents machine learning algorithm for analysis of medical data for determining health of hair. We use the SVM (support-vector machines) model classifier for analysis of data. After that, we get values such as short, straight, wavy and curly.In this paper, J48 Algorithms were used to obtain an accurate result compared with other algorithms.J48 with bagging is creating different decision trees for same data that why it is given more accurate results, J48 algorithms will split continuous values through using threshold. This paper 1066 samples were tested using cross validation technique, according to the test, it is found 87.14 % was a correctly classified and 12.85 % was incorrectly classifier. So at the end we get a real time performance is 89.5 %. This paper proves the compatible between hair style and Age-Gender.

Keywords: Spatial-temporal data; Hairstyle classification; Hair segmentation; Decision tree; J48 with bagging

Corresponding Author:

Ahmed Mahdi Abdulkadium,
Assist.lecture.Computer Science (information System),
Computer Center, Al-Qasim Green University,
60 Road, Al-Asatetha Region, Babylon, Iraq.
Email: ahmed_mahdi.@uoqasim.edu.iq

1. Introduction

Spatial-Temporal data is related to space and time. Many Data mining researches uses Spatial-Temporal data in their data analysis. Images captured from satellite cannot be analysed just by position of things, as this data dynamic in nature. For better analysis of this data, it is necessary to use images presenting in the different time as temperature and weather can be different at different time. Spatial is stand for space and Temporal stands for time variation. In Spatial-Temporal data, time and space images are used for analysis purpose [1].

Different Hairstyle leads to noisy data, this data can be analysed by using hairstyle detection system. It is found that Spatial-Temporal data for two different hair styles produce different analytical result. It proves that hair style of the specific person is also an important factor while determining hair score. It is basically to avoid false output, which get generated by long hair or short hair. Beside of long and short, hair can be straight, curly or wavy. Hair style detection further helps to find hidden pattern presented in the dataset.

Hair Images and Hairstyle are images dataset, Medical history is also important while analysis of hair growth. This includes age and gender of that person. Hair greying is common in aged people. Hair Greying is depending on age and gender of person. Hair loss statistic shows there is an effect of age and gender on hair

health. MPHL is a male pattern hair loss and FPHL is Female Pattern Hair Loss. In this statistic, it is observed that at age of 50, male losses half of hair, while female losses quarter of total hair [2].

Age and Gender are added as other attributes for knowledge discovery. Spatial-Temporal data, hair style analysis and medical history are some important aspects of hair health. Using data discovery technique, it is possible to find out hidden pattern present in the dataset.

There is some problems in collection of data. Spatial-Temporal data model is built by snapshot technique. It requires to study the scalp over different time, which may extend to five years. Snapshot of images is used to extract data. Dandruff is something which make picture unclear for analysis, which may lead to create noisy data. Oily scalp, which is affected of more moisture also affect quality of data. Cleaning is not enough capable solution to produce fine data. Spatial-Temporal data analysis requires performing analytical operations. These snapshots manly focus on hair follicle. Generally, hair follicle is capable of producing 3 hairs from one follicle, stem cell is a cause of hair greying and hair fall. Important factors from snapshot model is captured and divided into different group to use it as analytical parameters for data mining operations. This attribute is necessary to be patched to data before sending to machine learning algorithm. Age and Gender affects the analysis par too, it is observed that irrespective of age, few samples have great health, it may be because of hygiene, maintenance, food habits and less stressed over a time period. This factor makes changes in hair health of two different person of same age [3].

2. Data source and basic concept

2.1. Data source

Medical history of clients is collected over a time. Images of hair from scalp-scope microscope are recorded. This is then converted into frames; these images are taken at different time to get images at different time. Main focus of this images is on position, follicle and skin surface. Using these images follicles are detected. These data are collected over 4 to 5 years. Age and Gender of person are also collected which are required for decision tress as a parameter. Each Spatial-Temporal data with Age and Gender are considered as one entry. Along with that, hair style is captured from different angle. These is used as random attribute to decision tree.

2.2. Basic concepts

Follicle detection system: scalp-scope is used to detect the follicle present on the scalp. This follicle is analysed from hair points. Each Follicle contains 3 Hair points. These Hair points are responsible for hair growth. Using scalp-scope, follicle is detected and Hair count is measured from each follicle. This is a technique which most hair specialist preferred. Thickness of hair is also a measure parameter in considering hair health. Dino-Lite Microscope is used for capturing scalp images using 50x, 200x and 500x magnification.

Hair Style detection: There are already few techniques which are used for Face Detection, Gender recognition and video surveillance. This method basically analyses hair at patch level and segmented images at pixel level. After applying these methods, classification process is carried out, which classify hair tin to short, long, straight and curly field. This are some measure input to the decision tree as a parameter. These classified parameters are useful in the decision tree model formation process.

2.3. Data collection using spatial-temporal data model

As Explained in the previous section Spatial-Temporal model is based on Space and time images. These data should contain the spatial data denoting to various points which leads to follicle of hair. This spatial data also should be present in sequence which provides time stamping of the images, that it can be used as time series data. So that each follicle and its growth parameters related to time is stored in the databases for data analysis [4].

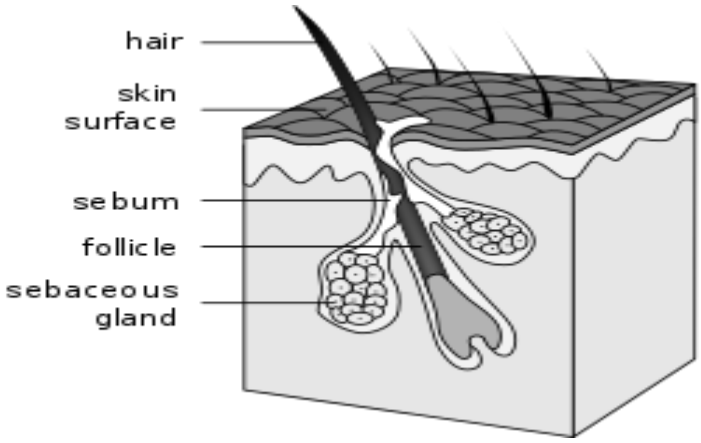

Figure 1. Hair Follicle

In this data model, each follicle is an act as a hair point and it contains 3 hair from one follicle. To maintain this information, structure data format is used. Root is name of structure holding all the information about that follicle and hair points. These roots object contains the information about data definition and analytical definition.

Data definition has fields for type of hair follicle, it is defined by sebaceous gland in medical terminology. It also has position of hair follicle and few functions which is about maintenance.

Another object which is **analytical definition**, contains analytical information which is strength of data, like how much the data points has importance. Follicle presented on the front side of head has less importance, while follicle presented at back side has more importance is deciding strength factor. Direction is one more field presented in analytical definition. These fields have knowledge of quantity and specifications of data. Like other systems which change with respect to time, these parameters also change with respect to time and the way person maintain his hair. Along with this analytical function are defined on natural specification of hair. Each function uses analytical and data definition of root object [3].

Before analysing current data set some pre-processings are required on the datasets. This is a procedure where raw data is converted into a dataset for machine learning [3]. There are few steps in pre-processing as below:

1) Cleaning: This is useful step because hair scalp may contain other particular which creates noise in image processing. Removal of such noisy data comes under cleaning step of data mining.
2) Integration: In this step different images are collected from different angles and the one which has more clear points and structure. This data is then integrated with strength factor.
3) Reduction: In data mining procedure some data can be there which is not required or which is repeated in the dataset, this data is removed by applying different data reduction technique, this involves Dimensionality reduction, aggregation and clustering.
4) Binarization: Captured image is coloured image which is converted to grayscale images. As because of direction and thickness of image, whole size image cannot be used, so all images are converted into 49*49 Pixels and then R function is used to convert it into grayscale image. Image processing library 'CImg' is used for conversion to grayscale image. This Library contains function 'grayscale' for conversion.

```
grayscale(im,method= "Luma",drop= TRUE)
```

In this Function im is a source of file. Method can be Luma or XYZ , Luma is for linear approximation , while 'XYZ' is used when an image is assumed to be as a RGB image. Drop parameter defined whether a return should be in single channel or not [5].

3. Methodology

3.1. Hair root extraction

Each hair point that is hair follicle detection is the main purpose for hair root extraction function. Lines are captured from 2 hair points; each line starts from hair points and ends at other end point ld. Then an imaginary line is drawn from that point with length le, in the opposite direction. If all other points are present on the line then it is not considered as a root. And the end, point is selected as a root. Using this technique hair root at different position are extracted and don't care hair root can be ignored as well as root with very thin layer and correctly extracted position.

Figure 2. Extracted positions

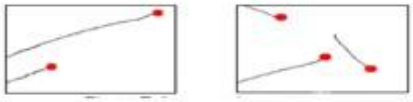

Figure 3. Enlarge sections

Figure 4 shows roots with different datapoints. Red points define the root which want to extract. Blue points define root with very thin hair. Yellow points denote wrongly extracted or don't care root. These data points are important for classification of strength of hair.

This technique is followed over a time, with the time stamp attached to it. Using this technique spatial-temporal data attribute is divided into strong, moderate and weak values [6][7].

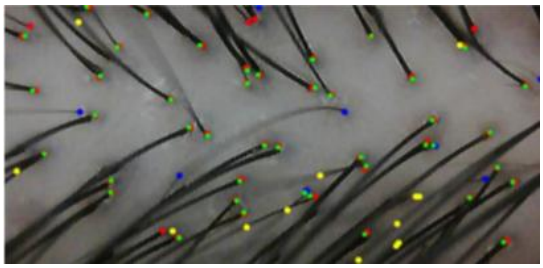

Figure 4. Extracted Hair Roots

3.2. Hair style classification

In Hair Analysis, many time miscalculations happen because of not considering hair style of person. Now a days many technologies have been evolved which uses hair style in face detection. similar technique is used to find out hair style of users. There is some problems in hair style analysis because of variation in hair style, dye and colour. Besides of this angle from which hair image is captured also make variation in view [8].

3.3. Hair detection on image patches

In the first step, probability map of presence of hair at patch level is captured. Hair texture detection system is used to apply classification pipeline for finding patches present in an image. In feature extraction phase, hair and non-hair patches are separated. because of this all the area containing no hair is eliminated. Texture descriptors are used to separate hair and non-hair region. In later stage, patched images are used for training model to create classifier. In the final stage, image is classified as segmented images where non-hair part has detected and excluded from analysis part. Different texture feature such as VGG-VD and LTP (Local Ternary pattern) is used for this detection [9].

$$\mathrm{LTP}_{P,R} = \sum_{p=0}^{P-1} 2^p s\left(i_p - i_c\right), \qquad s(x) = \begin{cases} 1, & x \geq t, \\ 0, & -t < x < t, \\ -1, & x < -t, \end{cases} \tag{1}$$

Where ic, ip defines grey value of centre pixel and from neighbour pixels inside the circle having Radius R and P is the total count of neighbours.

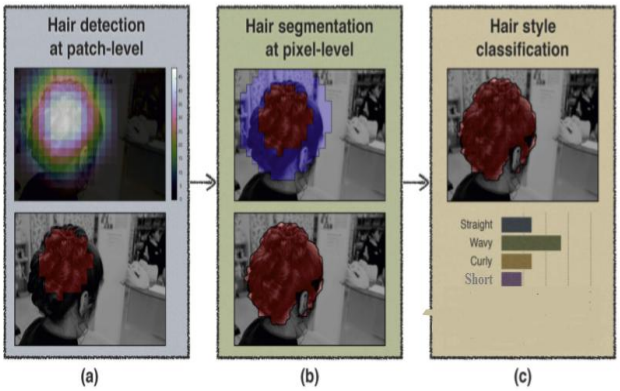

Figure 5. Hair Classification Model

Second step is Segmentation using pixels. In this step only, patches that are defined in the last phases are used. Segmentation is achieved by central pixel labelling scheme, in this method result of each patched classification determined in pervious step is assigned to central pixel. SVM classifier is used to classifier different instances of datasets [10].

$$\frac{1}{2} w^T w + C \sum_{i=1}^{N} \xi_i \tag{2}$$

This is an SVM model used for classification of data, where i…N are training cases, C is capacity constant, w stands for vector of coefficients, ξ_i is a parameter for considering nonsealable inputs. SVM have hyperplane to categorize given data into two or more classes. Suppose we are given with data set n in the form of

$$(\vec{x}_1, y_1), \ldots, (\vec{x}_n, y_n)$$

where y can be 1 or -1 depending on class of x.

In SVM, maximum-margin hyperplane is detected which divides all points of x into groups where y=1 and y= -1. This classification can be linear or nonlinear.

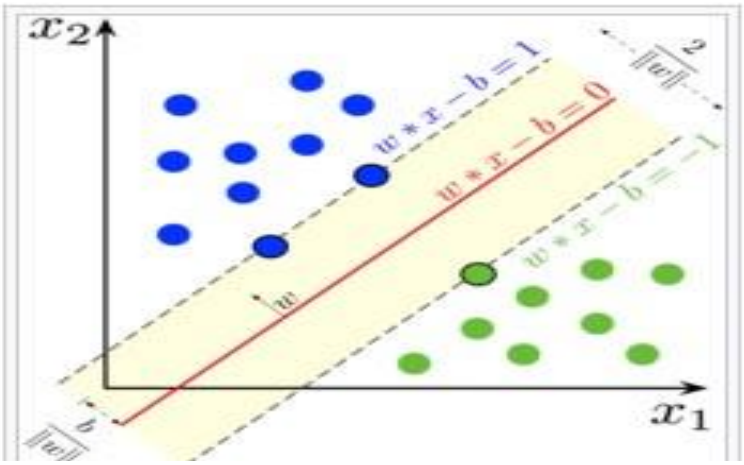

Figure 6. Linear SVM Classification

Hair Classification: Trained model, classifies new inputs to appropriate model these classifications have values such as short, straight, wavy and curly [2].

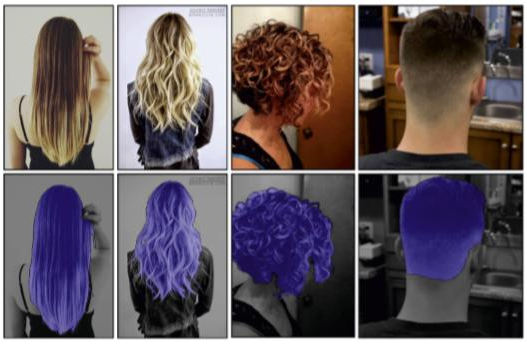

Figure 7. Hair Classification

The trained classification model performs multi-class classification on segmented images of hair. Voting schema is used for final labelling, label which has highest vote is selected [11].

3.4. Medical history

Medical history is also a linear attribute which was considered for classification of data. Medical history contains data like Age, Gender and hereditary. These factors affect the analysis of hair. It is observed that irrespective of age, few samples have great health, it may be because of hygiene, maintenance, food habits and less stressed over a time period. This factor makes changes in hair health of two different person of same age [12].

Age: Age is a something which really matters in concern of health check-up. It is observed that elder people are more prone to weak health and hair related issues too. Age is a parameter which is consider as linear numeric attribute for classification. Age is captured and stored in databases in the very early stage of analysis as it doesn't require pre-processing or other steps.

Gender: Gender alone is not so useful in hair analysis, but when gender and age are considered it is found to be co-relation between them. Males are more likely to lose their hair health early while in female this period is bit extended. Gender is also considered in analysis.

Hereditary: It's a simple measure with value either true or false. This contains hereditary information, meaning whether the person has any hereditary hair problem or not. If his parents or family person already has hair problem diagnostics then this value is set to true, or else it is set to false [13,14].

4. Model design

Data from this dataset is extracted. This information contains Spatial-Temporal Data, Hair style model, Age, Gender and Hereditary. Out of this Spatial-temporal data, it is divided into strong, moderate or weak parameters. This defines strength of hair. Hair style model which is described in previous stage results into Short, straight, wavy and curly label. Age is Numeric attribute which is fetched from the database. Gender is also extracted directly from Dataset. Gender can have value like Male and Female. Hereditary is another factor, with values either true or false.

This data is used for training purpose, using this training data, machine learning model is prepared. Different algorithm was trained to find the prefect training model which produces results with higher accuracy. It is observed that J48 with bagging has highest accuracy rate. Other classification algorithm has a problem of data overfitting, which is reduced in J48 algorithm by applying ensemble method. That is the reason, even some classification algorithms prove their accuracy higher, while working on real time scenario, their accuracy gets reduced. It is ensemble-based classification Algorithm. Generally, in classification algorithm, one classification tree is prepared to form machine learning model. In J48 with bagging algorithm value of n is need to specified, which is used to create a number of trees using random data set from whole data. This classification trees are constructed using different data set. Hence, it reduces the problem of data overfitting. After preparing n number of classification trees, voting technique is used to construct best classification model. Because of creating different decision trees, using the same data gives higher accuracy with respect to another algorithm. J48 is advanced decision tree in big data analysis [13]. Bagging is an ensemble technique. Very first step in constructing classification tree is finding the entropy rate of each class present in the dataset. Entropy rate of each class is calculated as below, and it is denoted as H(S).

$H(s) = \sum A(q) \log_2 (1 / A(q))$ (3)
where $s \varepsilon X$

In this equation s donates current data set, while q denotes classes present in s. After this Information gain of each class is calculated by,

$IG(A, S) = H(S) - \sum A(q) H(t)$ (4)
where $t \varepsilon T$

In this equation, t is a subset of S, H(t) is its entropy. A(q) is count of elements from in t with respect to s.

4.1. Improvement with respect to ID3 algorithm

- J48 can handle continuous and discrete class attribute. It split continues values by using threshold.
- It can handle missing attribute. Like "?"
- Data overfitting and Excess tree length issue can be handled by Tree pruning.
- J48 can support boosting and bagging technique.
- Speed for training and memory utilization is optimized as compare to ID3 (Iterative Dichotomiser 3).

4.2. Pseudocode for building J48 decision tree

- check for all base cases.
- Information gain of each attribute is calculated.
- select attribute with highest information gain.
- create decision tree on that node.
- keep this node as head and recur the same code for remaining attributes.

4.3. Bagging

Data overfitting issue of decision tree is handled by bagging technique. In this technique many bags of the same data are created which is later to find average. It is an ensemble technique. It keeps the bias same and reduces the variance. Multiple different bags are created which consists of random data selected from training set. These bags are later used as a training that in this way less variance is achieved. Bagging is also known as bootstrap aggregation. Bagging uses multiple copies of training data to create single model which is used for training purpose. J48 with bagging further reduces the problem of data overfitting by creating different class which provides better result than J48 alone [11].

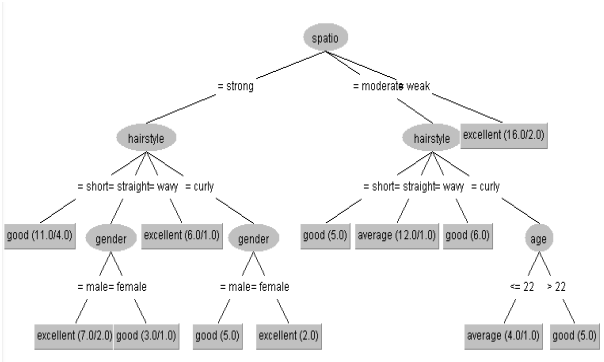

Figure 8. J48 Decision tree

J48 Decision tree can be created in weka with Rweka library. function to create machine learning model :

```
VarX <- Bagging(DEC~.,data=r,control =
                Weka_control(W="J48"))        (3)
```

where r stands for training data.

4.4. Machine learning model

Training model uses pre-processed data as an input. j48 algorithm performs machine learning algorithm to develop machine learning model. This Model takes input and produces output with excellent, Good, average and damaged parameters.

As Explained in Figure, pre-proceed data from different sources like Spatial-Temporal Data, Hair style model, Age, Gender and Hereditary are collected.

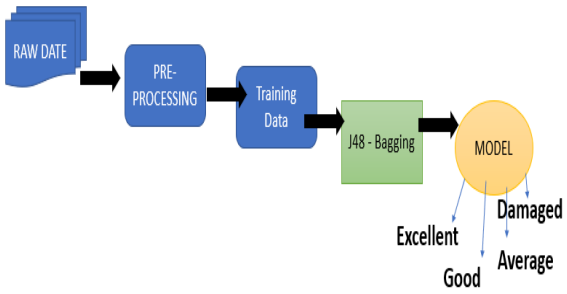

Figure 9. Hair Analysis using J48

New input is provided to j48 with bagging machine learning model. As result of analysis, the result predicts the output from specified parameter with the prediction rate. Bagging is an ensemble method and hence it requires little bit more time for model creation and analysis [12].

5. Experimental results

Accuracy of this algorithm is determined by confusion matrix. True positive rate and false positive rate is determined for performance analysis. Weighted true positive rate is 0.871 and false positive rate is 0.05. Precision and recall are also important while analysing performance of machine learning model. Recall is a measurement of truly classified instances. Recall is also known as sensitivity.

	TP Rate	FP Rate	Precision	Recall	F-Measure	Class
	0.934	0.080	0.859	0.934	0.895	excellent
	0.983	0.036	0.943	0.983	0.962	good
	1.000	0.029	0.875	1.000	0.933	average
	0.094	0.033	0.262	0.094	0.138	damaged
Avg.	0.871	0.050	0.828	0.871	0.844	

For performance analysis 10-fold cross validation is used. It uses 10 iteration with different training and testing data. Cross-validation is well known analysis technique. In this paper, 1066 sample instances were trained and tested using cross validation technique. In this analysis, it is found that correctly classified Instances are 87.14% and incorrectly classified instances are 12.85%. Root mean squared error are 0.1884.

Table 1. J48 Classifications

Summary	Inst	Percentage
CorrectlyClassified Instances	929	87.142 %
incorrectlyClassified Instances	137	12.8518 %
Root Mean Square Error	0.118	-
Relative Absolute error	-	21.03 %
Total Instances	1066	-

While working with ensemble trees, confusion matrix is not enough. That why out of bag estimation is also important, Weka tool is used to visualize the OOB (out-of-bag)error of by excellent, good, damaged and average classes.

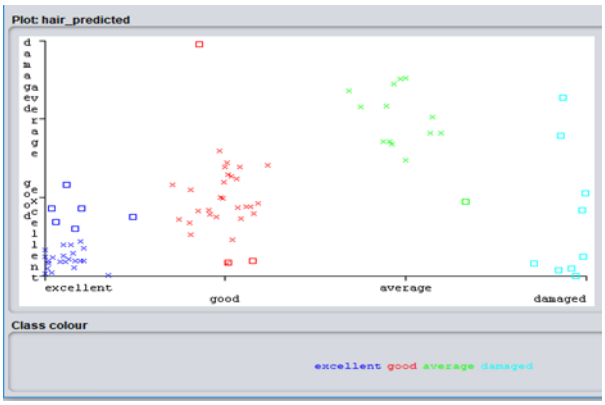

Figure 10. Experimental result

For Real time analysis, 1500 samples were tested, out of which output probability distribution of 0.6 or more was selected, it is found that out of 1150 instances, 1035 are correctly classified. Detailed table with all instances is present below.

```
=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 340   0   0  24 |   a = excellent
   0 396   0   7 |   b = good
   0   0 182   0 |   c = average
  56  24  26  11 |   d = damaged
```

It is observed that real time accuracy of model is 89.5% while n-fold cross validation accuracy of model is around 87.14 %. It proves that this model has balanced data and real time accuracy justify the particle use of this model in Hair analysis research.

6. Conclusion

6.1. Findings

In this paper, J48 with bagging is used to create machine learning model. This model uses ensemble method which creates multiple decision trees. Using this method effect of data overfitting is handled. It is found that accuracy of this model is 87.14 %. This model also proves the importance of hair style and Age-Gender analysis in hair health detection. Real time performance of this model is 89.5%.

6.2. Further research

Different data handling techniques like neural network and clustering can be used. Different parameters can be added. More attributes like hormone level and other existing disease can be added, these steps can be implemented to improve performance.

Acknowledgments

This work was supported from Aredo Center Beauty and Karim Murad Salon through providing a sample of hair. These samples were analyzed and considered the important criteria in the analysis of hair characteristics, with the help of Ibn Al Haytham Center for Pathological Analysis through classification of hair types under the microscope where the hair was classified such as wavy, tape and curly hair.

References

[1] M. Abbas, A. O. Zuleika, H. Abdul Razack. ”Hair Data Model: A New Data Model for Spatial-Temporal Data Mining”, 4th Conference on Data Mining and Optimization (DMO), 2012.

[2] N. J. Qasim and I. Barazanchi, "Unconstrained Joint Face Detection and Recognition in Video Surveillance System," *Jour Adv Res. Dyn. Control Syst.*, vol. 11, no. 1, pp. 1855–1862, 2019.
[3] N. Shun , T. Masanobu." Image analysis of hair - Hair roots extraction - Proceedings of the SICE ",Annual Conference , Kanazawa University, Kanazawa, Japan, IEEE access,2017. , Available from http://ieeexplore.ieee.org/document/8105560.
[4] J. J. Seong , H. P. Seung , W. C. Jae, H. L. Jong , C. Soyun, H. K. Kyu, H. C. Hee and S. K. Oh ."Hair Graying Pattern Depends on Gender, Onset Age and Smoking Habits", 92(2),pp.160-161,2016.
[5] A. Parham, "Automatic segmentation of hair in images", 2015 IEEE International Symposium on Multimedia (ISM), IEEE Access, 2015. Available from
https://ieeexplore.ieee.org/document/7442299.
[6] H. G. Zhang, and M. Piccardi. "An accurate algorithm for head detection based on XYZ and HSV hair and skin color models",15th IEEE International Conference on Image Processing. , IEEE Access, 2008. Available from: https://ieeexplore.ieee.org/document/4712087.
[7] Y. Yacoob, L.S. Davis, " Detection and analysis of hair", IEEE Trans, Pattern Anal. Mach. Intel, 28 (7) ,pp.1164–1169,2006.
[8] U. Toseeb, D.R. Keeble and E.J. Bryant. "The significance of hair for face recognition", [PloS one], 7 (3) e34144, 2012.
[9] M. Chai, T. Shao, H. Wu, Y. Weng and K. Zhou , "Auto hair: fully automatic hair modelling from a single image", [ACM Transactions on Graph], 35 (4),pp.116,2016.
[10] Y. Wang, Z. Zhou, E.K. Teoh and B. Su, "Human hair segmentation and length detection for human appearance model", presented at the conference,22nd International Conference on Pattern Recognition (ICPR),2014. IEEE Access, Available from https://ieeexplore.ieee.org/document/6976797.
[11] K. R. Pradeep and N. C. Naveen. "Predictive analysis of diabetes using J48 algorithm of classification techniques",presented at the conference,2nd International Conference on Contemporary Computing and Informatics (IC3I) , 2016.IEEE Access, Available from https://ieeexplore.ieee.org/document/7917987.
[12] H. Indriana, E.P. Adhistya and A. K. Monica . "Application of J48 and bagging for classification of vertebral column pathologies",Presented at the conference(Proceedings of the 6th International Conference on Information Technology and Multimedia) ,2014.IEEE Access , Available from https://ieeexplore.ieee.org/document/7066651.
[13] N. Juthamas, S. Supaporn . "Kiattisin and Adisorn Leelasantitham.Diagnosis and interpretation of dental X-ray in case of deciduous tooth extraction decision in children using active contour model and J48 tree" , International Electrical Engineering Congress (IEECON),2014. IEEE Access , Available from https://ieeexplore.ieee.org/document/6925902 .
[14] S. Rashid, A. Ahmed, I. Al Barazanchi, and Z. A. Jaaz, "Clustering algorithms subjected to K-mean and gaussian mixture model on multidimensional data set," *Period. Eng. Nat. Sci.*, vol. 7, no. 2, pp. 448–457, 2019.

BIBLIOGRAPHY OF AUTHORS

Ahmed Mahdi Abdulkadium is a lecturer at Al_Qasim Green University , Iraq. He received the Bsc.degree in computer Science from University of Babylon,2012, the MSc. Degree in Information System from Nizam College ,college of science , Osmania University in 2015,his current research interest include computer science , information system development, data mining, data warehouse and machine learning.
ahmed_mahdi.@uoqasim.edu.iq

Raid Abd Alreda Shekan is a lecturer at Babylon University , Iraq. He received the Bsc.degree in computer Science from University of Babylon,1995, the MSc. Degree in Information System from Nizam College ,college of science , Osmania University in 2015,his current research interest include computer science , information system development, data mining, Knowledge development . pure.raed.abd.@uobabylon.edu.iq.

Ali Abdulbaqi Abdulazeez is a teacher at Ministry of education/General Directorate of Basrah Education , Iraq. He received the Bsc.degree in computer Science from University of Basra,2008, the MSc. Degree in Information System from Nizam College ,college of science , Osmania University in 2015,his current research interest include data analysis , information system development, data mining, Knowledge development.
ali.alfahad.2006@gmail.com

.