Evaluation Machine-Learning Approaches for Classification of Cryotherapy and Immunotherapy Datasets

Ali CÜvitoğlu and Zerrin Işik

***Abstract*—Machine-learning (ML) methods have great importance when applied interdisciplinary. Besides many areas, ML methods save cost and time in medical applications. In this study, we experimented several ML methods with different approaches on classification of Cryotherapy and Immunotherapy datasets, which are applied on wart treatment. The effects of dimension reduction techniques and handling of unbalanced sample classes are the main discussion points of our study. When several ML models are analyzed, Random Forest (RF) achieved 95% accuracy, %88 sensitivity, and %98 specificity. Other ML methods also performed successful results close to the RF. Although some promising results were obtained, we also discussed the drawbacks of these approaches while evaluating wart treatment strategies.**

***Index Terms*—Machine-learning methods, principal component analysis, linear discriminant analysis, cryotherapy, immunotherapy, wart treatment.**

I. Introduction

Nowadays, machine-learning (ML) methods are applied in many different medical applications such as understanding the disease developments, diagnosing and choosing a treatment method. Recently, a fuzzy logic rule-based system was proposed and implemented to predict if warts will be healed by the wart treatment methods such as cryotherapy and immunotherapy [1]. In this study, a cohort for 180 patients has been collected by the dermatology clinic of Ghaem Hospital, Mashhad, Iran. Ninety of these patients were treated with immunotherapy method and other ninety patients by cryotherapy method to get rid of plantar and common warts. The aim of that study was to predict the response of the treatment to select a right therapy due to many sessions are required for healing warts. For these reasons, Khozeimeh F *et al.* compared a classic rule based and a fuzzy rule-based method as classifiers.

There is more than one wart treatment method such as electrocautery, surgical removal, laser ablation, intralesional injection of bleomycin, Candida albicans (C. albicans), purified protein derivatives (PPD), and mumps, measles, rubella (MMR) antigens [2], [3]. However, all methods may have different side effects or other difficulties. Warts also can be infectious, so they must be treated at the same time, extra treatment will take time due to number of sessions. One of wart treatment methods, Immunotherapy has also been used for the treatment in children. Clifton *et al.* treated 47 children with intralesional injection of mumps or Candida skin test antigen [4]. 47% of patients are completely treated and 34% of the children showed more than 25% healing in their warts. Silverberg *et al.* experimented squaric acid dibutylester on 61 children [5]. In this study, 58% of children completely treated while %18 partially treated. Another study reported their experiences with intralesional candida antigen therapy [6]. Here, warts are cleared completely for 56% of 217 patients and %28 treated partially. Khozeimeh F *et al*. proposed another study to compute the efficacy of immunotherapy and cryotherapy on wart lesions [3]. In that study, an immunotherapy method was applied three weeks and cryotherapy was applied ten weeks to patients. Although they found that immunotherapy was more effective as therapeutic response, similar success rates were generally observed for both methods.

In this study, we received two datasets published by Khozeimeh *et al.* [1], [3]. These datasets contain information on whether Cryotherapy and Immunotherapy treatment methods are successful when applied for wart treatment. We aimed to study on these datasets since warts might require long term treatments. Besides, some type of warts can be infectious. Due to these facts, choosing an appropriate treatment is a critical issue for this disease. We tried to come up with several approaches of machine learning methods to choose an appropriate wart treatment method. In accordance with this purpose, different standard ML methods such as Naive Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), k Nearest Neighborhood (kNN), and Artificial Neural Network (ANN) experimented and compared with each other on these datasets. Dimensionality reduction and feature selection approaches are applied to measure whether the success rates will increase. As another approach, we aimed to run ordinal classification [7]. Because we hypothesis that there is a sequential relationship between the labels (Result of Treatment; Yes > No) as we will discuss in the next sections.

II. Method

A. System Overview

In this study, we obtained two datasets, which cover cryotherapy and immunotherapy treatment details. We applied different approaches on the datasets. As seen in Fig. 1, our model contains four parts. First, it is optional to choose, all features are used as input or applied either Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA). The second step selects 90% of dataset for training and builds ML methods with these

Manuscript received April 11, 2018; revised June 26, 2018.

The authors are with the Deparment of Computer Engineering Department, Dokuz Eylül University, Izmir, 35370 Turkey (e-mail: ali.cuvitoglu@cs.deu.edu.tr, zerrin@cs.deu.edu.tr).

data. In the third step, remaining 10% of data is used to evaluate ML methods and the final step compares different evaluation methods. This is shown for one-fold Cross-Validation (CV) with one dataset. 10-fold CV for both datasets is computed for each ML methods.

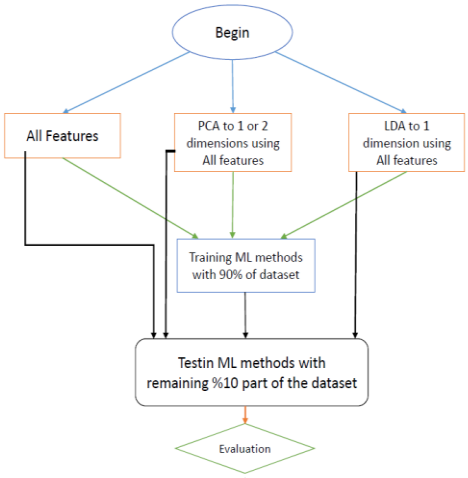


Fig. 1. The pipeline of the proposed model.

B. Datasets

We received Cryotherapy and Immunotherapy datasets from UCI database [8] published by [1], [3]. Each dataset contains records for ninety patients. Cryotherapy and Immunotherapy datasets consist of 6 and 7 features, respectively (Table I).

TABLE I: FEATURES OF DATASETS

	Values	
Features	**Cryotherapy**	**Immunothreapy**
Age	15-67	15-56
Gender (M-F)	47-43	41-49
Time	0-12	0-12
Number of Warts	1 -- 12	1 -- 19
Area of Warts (mm^2)	4 -750	6-900
Type	Common /Plantar /both	Common /Plantar /both
Induration Diameter	--	5 -- 70
Result of Treatment	Yes or No	Yes or No

Age of patients starts from 15 and increases. Almost the same proportion of female and male patients is treated by each treatment. Time elapsed before treatment is between 0 and 12. Number of warts are mostly 12 and 19 for Cryotherapy and Immunotherapy, respectively. Area of the biggest warts and type of these warts are also included. The difference between datasets is Cryotherapy doesn't contain induration diameter of the initial test. Thus, the term of 'all features' cryotherapy and for immunotherapy covers in total 6 and 7 features, respectively. The labels of these datasets are given "Yes" (positive) or "No" (negative). If the size of the biggest wart decreased by >75%, it is considered positive; it will be negative, if it is less than 25%. If the reduction was between 25% and 75%, this is also considered as negative by the publisher of the datasets [3]. If this part of the datasets would have been published, it would be a better approach for ordinal classification (OC) with three labels like A (>75%) > B (75%> and >25%) > C (<25%). In the current version, for an unseen sample, OC would compute *Pr(sample > C), Pr(sample > B)* where *Pr* is the probability of being in higher ranked classes (e.g. A) of the given class (B). Then, the probability of being in class A, B, and C is *Pr(sample > B), Pr(sample > C)x(1-Pr(sample > B))* and *1-Pr(sample > C)*, respectively. The sample will attend to the class with the highest probability [7]. However, OC will resemble NB in binary classification.

C. Data Preprocessing, Principal Component Analysis and Linear Discriminant Analysis

Data preprocessing is one of the most critical stage of a ML study. For example, feature selection helps us to choose features with a high discrimination power. *T*-test is applied with threshold *p*-value <= 0.05. Three (Age, Time, Type) and two features (Time, Type) passed the test for Cryotherapy and Immunotherapy datasets, respectively. However, these features alone may not be meaningful for the classification. Cryotherapy dataset can be considered as balanced due to having 48 positive and 42 negative samples. However, immunotherapy dataset has 71 positive and 19 negative samples which is unbalanced. 'Smote' is an oversampling method can be used to create synthetic data to balance the dataset. We installed DMwR package in R to apply the smote method on Immunotherapy dataset.

Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) are applied to decrease dimensions of the feature space. Ninety samples for 6 or 7 features are very small. PCA uses eigenvectors of the covariance matrix to identify independent axes of the data under the unimodal Gaussian assumption, whereas LDA finds a linear combination of features that separates two or more classes. Using 'factoextra' package in R, we projected features to only 1 or 2 dimensions for PCA approach and 'MASS' package for LDA approach. In this package, 'lda' function decides the dimension according to number of labels. If there are *n* labels, it gives *n*-1 dimensions. The data contains two labels, so we obtained one dimension for LDA.

D. Usage of Machine-Learning Methods

Various types of machine-learning methods were executed. These ML methods are Naïve Bayes (NB), Random Forest (RF), Support Vector Machine (SVM), Artificial Neural Networks (ANN) and k-Nearest Neighbor (kNN) algorithm. Additionally, there is an ensemble learner, which is the combination of these five algorithms. It applies a typical voting system for ML methods. For each sample predicted by ML methods, if at least three of them predict as positive, the final class of this sample will be positive by the ensemble learner, otherwise it will be negative. Another method is Ordinal Classification (OC) which is a probabilistic approach [4]. Here, we assumed *positive > negative* as order. For unseen sample, it computes it's probability of being higher than a negative: *Pr(sample > negative)*. Hence, *1-Pr(sample > negative)* gives the probability of being negative. Here, a higher probability value determines the class of the sample. Although OC is better to apply for multiclass classification [7], we aimed to test it for binary classification.

For kNN, the 'class' package is used, and *k* value was determined by experimenting values from 1 to 55 increasing by two. Finally, *k* was set to 1. For ANN, we utilized package called 'neuralnet'. The X-3-1 structures were designed where X is the input size (feature size). The

learning rate was set to 0.01. The backpropagation algorithm was selected as the learning strategy. For RF, we used a package called 'randomForest' and the parameter for trees is set to 50. For NB and SVM, we utilized 'e1071' package in R. For SVM, this package offers a function to predict the cost and gamma values; radial kernel function is applied. For OC, 'ordinal' package provides 'clm' function.

E. Evaluation

Application of a Cross-Validation (CV) scheme has a significant impact on the evaluation of a ML method. Ninety samples partitioned to 10. For each part to be test set, 10-fold CV is completed, and the average performance of 10-fold is computed.

As evaluation method, we measured accuracy, precision, recall, and F1-measure that can be calculated from the confusion matrix. Additionally, Receiver Operator Characteristic (ROC) curve and Area under the ROC curve (AUC) are computed.

Accuracy gives the percentage of correctly estimated results. Classification algorithms are designed to classify all classes correctly. In this case, as well as the true positive (TP), the performance of the system is affected by the true negative (TN).

Precision is the measure of certainty or quality while recall is the measure of completeness or quantity. Precision value indicates true positive rate of all positive predictions while recall represents true positive rate of all actual positive samples.

F1-measure is a harmonic mean of precision and recall. This measure can be considered as an average of these two measurements.

ROC Curve is used to visualize the performance of a classifier using more than one threshold. The AUC gives a specific number as the summary of the curve. ROC uses True Positive Rate (TPR) and False Positive Rate (FPR). TPR is known as sensitivity and FPR is known as *1-specificity*. For each threshold, the confusion matrix is computed, TPR and FPR can be calculated from the confusion matrix. Every TPR and FPR generates a point of the curve. After all thresholds, all points create a curve that is called ROC curve.

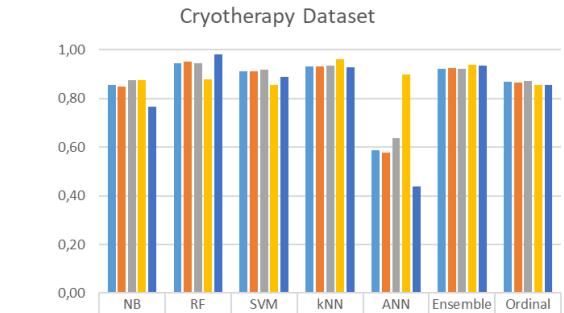
III. Results

There have been several runs in this study. The average performance of 10-fold CV for each ML method is computed by applying each evaluation. Here, we will take a glance at results in four sections.

A. Comparison of Results as All Features Input

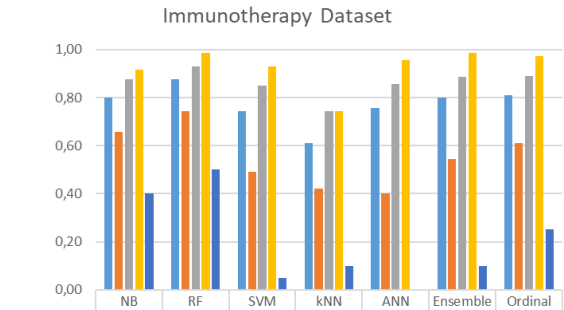
Six different types of standard ML methods have been experimented with also ensemble learner in this analysis. All features are given as the input of each ML method. We applied 10-fold CV and the results contain the average performance of 10-fold CV while parameters of ML methods have been set as explained in the Methods-section D.

Fig. 2 and Fig. 3 show the results for Cryotherapy and Immunotherapy dataset, respectively. The 'Smote' method is not applied for Immunotherapy dataset for these computations.



	NB	RF	SVM	kNN	ANN	Ensemble	Ordinal
Accuracy	0,86	0,95	0,91	0,93	0,59	0,92	0,87
AUC	0,85	0,95	0,91	0,93	0,58	0,93	0,87
F1-Measure	0,88	0,94	0,92	0,94	0,64	0,92	0,87
Sensitivity	0,88	0,88	0,86	0,96	0,90	0,94	0,86
Specificity	0,77	0,98	0,89	0,93	0,44	0,94	0,86

Fig. 2. Results of Cryotherapy dataset using all features as the input of ML methods.



	NB	RF	SVM	kNN	ANN	Ensemble	Ordinal
Accuracy	0,80	0,88	0,74	0,61	0,76	0,80	0,81
AUC	0,66	0,74	0,49	0,42	0,40	0,54	0,61
F1-Measure	0,88	0,93	0,85	0,74	0,86	0,89	0,89
Sensitivity	0,92	0,99	0,93	0,74	0,96	0,99	0,97
Specificity	0,40	0,50	0,05	0,10	0,00	0,10	0,25

Fig. 3. Results of Immunotherapy dataset using all features as the input of ML methods.

When the results of Cryotherapy dataset (Fig. 2) are analyzed, three methods and the ensemble learner have passed 90% accuracy, RF has the highest performance. NB and OC might also be considered successful with 86% and 87% accuracies, respectively. However, ANN failed to achieve even 60% accuracy. We tested ANN with different parameters, however, we couldn't obtain better results with ANN. The other evaluation metrics like sensitivity and specificity, shown in Fig. 2, that provided between 86% and 98% except NB and ANN. AUC value can be considered successful, when it is higher 0.9. So, RF, SVM, kNN, and the ensemble learner achieved successful resGults based on AUC values.

Fig. 3 shows the results of Immunotherapy dataset. Here, RF is again the best performing ML method, by far. The reason of poor results of other ML methods might be unbalanced data set when compared to Cryotherapy one. 71 positive and 19 negative samples were available in this dataset. Sensitivity and specificity results show that while positive samples were classified correctly, negative samples were also classified as positive ones. So, the models cannot discriminate negative samples from positive ones, effectively.

B. Comparison of PCA and LDA

PCA and LDA are commonly used when there are not enough samples to represent the feature space. Six features create a six dimension. Ninety samples data set is very low

number to represent six dimensions in an appropriate feature space. Here, we applied PCA and LDA approaches to decrease six dimensions to one dimension. Results cover the average of 10-fold CV. Table II and Table III show the results of PCA and LDA for Cryotherapy and Immunotherapy datasets, respectively. When both datasets and ML methods are considered, LDA outperformed PCA. LDA also has more consistency based on sensitivity and specificity metrics compared to the PCA. The difference can be seen in F1-measure and AUC results, as well.

The main reason to apply a dimensionality reduction is to increase classification performances compared to usage of entire feature space. When LDA and PCA results compared with results in the Fig. 2 and Fig. 3, ANN results draw the attention. Due to restrictions in complexity and run time, it was not possible to try higher number of neurons in the hidden layer of ANN. Therefore, the performance of ANN increased when LDA and PCA are applied.

TABLE II: PCA VS LDA ON THE CRYOTHERAPY DATASET BY USING ONLY ONE DIMENSION

Cryotherapy dataset					
PCA		**NB**	**RF**	**SVM**	**ANN**
	Accuracy	0,82	0,79	0,81	0,80
	AUC	0,82	0,79	0,81	0,92
	F1 - Measure	0,85	0,80	0,82	0,83
	Sensitivity	0,92	0,82	0,86	0,92
	Specificity	0,72	0,77	0,76	0,67
LDA		**NB**	**RF**	**SVM**	**ANN**
	Accuracy	0,90	0,86	0,89	0,89
	AUC	0,91	0,86	0,89	0,97
	F1 - Measure	0,90	0,86	0,89	0,89
	Sensitivity	0,86	0,86	0,86	0,86
	Specificity	0,96	0,86	0,93	0,93

TABLE III: PCA VS LDA ON THE IMMUNOTHERAPY DATASET BY USING ONLY ONE DIMENSION

Immunotherapy Dataset					
PCA		**NB**	**RF**	**SVM**	**ANN**
	Accuracy	0,79	0,67	0,79	0,80
	AUC	0,50	0,53	0,52	0,69
	F1 - Measure	0,88	0,78	0,88	0,89
	Sensitivity	1,00	0,76	0,99	0,99
	Specificity	0,00	0,30	0,05	0,10
LDA		**NB**	**RF**	**SVM**	**ANN**
	Accuracy	0,84	0,74	0,82	0,83
	AUC	0,69	0,64	0,69	0,78
	F1 - Measure	0,91	0,83	0,89	0,90
	Sensitivity	0,97	0,83	0,93	0,93
	Specificity	0,40	0,45	0,45	0,50

NB also improved the accuracy up to 90% in LDA on the Cryotherapy dataset.

C. Effect of Oversampling on Immunotherapy Dataset

In this part, we used 'smote' function in 'DMwR' package in R for oversampling the negative samples of Immunotherapy dataset. The negative samples are quadrupled. Thus, 71 positive samples and 76 negative samples were used with all features.

When Fig. 3 and Fig. 4 are compared, the specificity results are clearly improved while sensitivities are decreased. Even though RF is decreased 2% after applying the smote method, it can be accepted more successful due to balanced sensitivity and specificity results. AUC and F1-measure also show higher values after applying the smote method.

D. Comparison with Previous Studies

Khozeimeh *et al.* [1] proposed a model to classify cryotherapy and immunotherapy datasets. The model is a fuzzy rule-based method that uses adaptive network-based fuzzy inference system [9] to optimize membership functions. Their method was compared with the classic rule-based method. The results of their method are shown in Table IV and Table V listed with our top selected results.

Cryotherapy dataset is a balanced dataset. Even though there are few samples, standard ML methods achieved higher accuracy results between 86% and 95%, which outnumbered the classic and fuzzy rule-based methods. Sensitivity and specificity results also provide true predictions of both classes with higher percentages than the results of the previous study.

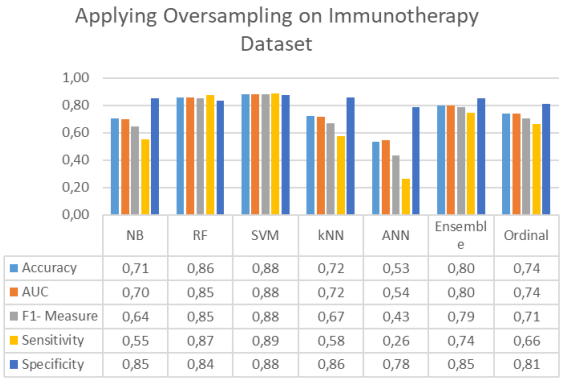


Fig. 4. Results of ML methods when oversampling was applied on Immunotherapy dataset.

TABLE IV: COMPARISON OF RESULTS WITH PREVIOUS STUDY [1] WHEN CRYOTHERAPY DATASET IS USED

Cryotherapy		Accuracy	Sensitivity	Specificity
All Features	**NB**	0,86	0,88	0,77
	RF	0,95	0,88	0,98
	SVM	0,91	0,86	0,89
	kNN	0,93	0,96	0,93
	Ensemble	0,92	0,94	0,94
	Ordinal	0,87	0,86	0,86
LDA	**ANN**	0,89	0,86	0.93
Classic Rule Based Method [1]		0,7	0,7	0,69
Fuzzy Rule Based Method [1]		0,8	0,82	0,77

TABLE V: COMPARISON OF RESULTS WITH THE PREVIOUS STUDY [1] WHEN IMMUNOTHERAPY DATASET IS USED

Immunotherapy		Accuracy	Sensitivity	Specificity
After Oversampling	**RF**	0,86	0,87	0,84
	SVM	0,88	0,89	0,88
LDA	**ANN**	0,83	0,93	0,5
Classic Rule Based Method [1]		0,73	0,78	0,57
Fuzzy Rule Based Method [1]		0,83	0,87	0,71

Using different ML approaches in classification problems

may lead better predictions. However, raw data must be analyzed before deciding the ML method. Here, we experimented standard ML methods on two datasets. Some of ML methods achieved better results compared to the own method of the dataset's publisher. However, fuzzy rule-based method [1] closed the gap between sensitivity and specificity. Classic ML methods may fall into error when dataset is unbalanced. Our experiments showed that when a data set has unbalanced class samples, the samples from outnumbering class is generally classified more efficiently. We managed to close this gap by applying oversampling technique. After oversampling, SVM and RF performed higher accuracy values than the results of previous study [1]. ANN also managed to have high accuracy. However, ANN classified more positive samples than negative ones when only LDA is applied.

IV. Conclusion and Future Work

In this study, newly published datasets were analyzed for wart treatment classification purpose by applying several ML methods. We obtained promising results by applying and comparing different techniques. Every technique has both advantages and disadvantages depend on the problem.

Cryotherapy and Immunotherapy are significant wart treatment methods. Conventional classification approaches utilized to decide whether the treatment would be remedial based on given features. The publisher of the datasets was implemented a new model for this problem due to the necessity of new models for such problems. However, the small number of samples and unbalanced classes were not considered in their study. We tried to solve unbalanced samples between two classes by applying an oversampling on the samples in the lower numbered class. Although synthetically produced samples may not represent a real sample for a patient, the classification performance slightly improved in some evaluation metrics.

In the future work of these studies, we believe if the target class number is incremented to three classes as explained in the section Methods–B, OC method might provide better and more meaningful predictions. Although OC is a multiclass classification method, its performances in our study are promising even for the binary classification. Therefore, an improvement of the proposed classification models can be accomplished by using a multiclass OC.

References

[1] F. Khozeimeh, R. Alizadehsani, M. Roshanzamir, A. Khosravi, P. Layegh, and S. Nahavandi, "An expert system for selecting wart treatment method," *Computers in Biology and Medicine*, vol. 81, pp. 167-175, 2017.

[2] S. Gibbs, I. Harvey, J. Sterling *et al*., "Local treatments for cutaneous warts: systematic review," *BMJ*, vol. 325, no. 461, 2002.

[3] F. Khozeimeh *et al*., "Intralesional immunotherapy compared to cryotherapy in the treatment of warts," *International Journal of Dermatology*, vol. 56, no. 4, pp. 474-478, 2017.

[4] M. M. Clifton, S. M. Johnson *et al*., "Immunotherapy for recalcitrant warts in children using intralesional mumps or Candida antigens," *Pediatr Dermatol*, vol. 20, pp. 268–271, 2003.

[5] N. B. Silverberg *et al*., "Squaric acid immunotherapy for warts in children," *J Am Acad Dermatol*, vol. 42, pp. 803–808, 2000.

[6] M. Maronn, C. Salm *et al*., "One-year experience with candida antigen immunotherapy for warts and molluscum," *Pediatr Dermatol*, vol. 25, pp. 189–192, 2008.

[7] E. Frank and M. Hall, "A simple approach to ordinal classification," *Lecture Notes in Computer Science*, vol. 2167, 2001.

[8] UCI KDD Archive. [Online]. Available: http://archive.ics.uci.edu/ml/

[9] J.-S. R. Jang and C.-T. Sun, "Neuro-fuzzy and soft computing: A computational approach to learning and machine intelligence," Prentice- Hall, Inc, 1997.

Ali Cüvitoğlu was born in Hatay, Turkey on August 14, 1992. Cuvitoglu is a PhD student in Computer Engineering Department, Dokuz Eylul University (DEU). Cuvitoglu studied a bachelor in Cukurova University, then finished his master in DEU. He is now a research assistant in DEU.

Zerrin Işık got a Ph.D. degree from Computer Engineering Department of Middle East Technical University in 2011. Her Ph.D. dissertation established a novel pathway enrichment system based on integration of gene expression, ChIP-sequencing data and cyclic signaling pathways to assess biological activity of specific cell processes. She worked as a post-doctoral researcher in Biotechnology Center of TU Dresden, Germany from 2011 to 2014. She works as an assistant professor in the Department of Computer Engineering of Dokuz Eylül University since 2014.