

# Kidney X-ray Images Classification using Machine Learning and Deep Learning Methods

Işıl Karabey Aksakallı, Sibel Kaçdioğlu and Yusuf Sinan Hanay

**Abstract**—Today, kidney stone detection is performed manually by humans on medical images. This process is time-consuming and subjective as it depends on the physician. This study aims to classify healthy or patient individuals according to the status of kidney stones from medical images using various machine learning methods and Convolutional Neural Network (CNN). We evaluated various machine learning methods such as Decision Trees (DT), Random Forest (RF), Support Vector Machines (SVM), Multilayer Perceptron (MLP), K-Nearest Neighbor (kNN), Naive Bayes (BernoulliNB), and deep neural networks using CNN. According to the experiments, the Decision Tree Classifier (DT) has the best classification result. This method has the highest F1 score rate with a success rate of 85.3% using the S+U sampling method. The experimental results show that the Decision Tree Classifier (DT) is a feasible method for distinguishing the kidney x-ray images.

**Index Terms**—kidney disease detection, classification, deep learning, machine learning, artificial intelligence in medicine.

## I. INTRODUCTION

**K**IDNEY STONE disease occurs due to the accumulation of salt and mineral crystals excreted in the urine and turning into stones. Kidney stones have been affecting people for centuries. It is one of the most common diseases of the kidneys and urinary tract. It affects approximately 1-15% of the world population, and its prevalence is increasing with each passing year [1]. The prevalence of kidney stones was 1–5% in Asia, 5–9% in Europe, and 7–15% in North America [1]. Kidney stone formation can occur at any age, gender, and


race. It is more common in men than in women. Kidney stones are thought to be caused by reasons such as lack of physical activity and eating habits. Chronic diseases such as blood pressure, diabetes, and obesity may affect stone formation. After the kidney stone is treated, it may recur and become chronic.

Prevention of kidney stone formation and recurrence is still a significant problem for human health. Impairment of kidney function due to the formation of kidney stones endangers human life. Therefore, early diagnosis of kidney stones is critical. In recent years, machine learning and deep learning approaches have been widely adopted to diagnose diseases thanks to the development of technology. These methods provide a reliable tool for making definitive diagnostic decisions that require long and complex processes, as they shorten the diagnosis time and increase the diagnostic accuracy. Along with deep learning, computer vision can categorize the images, extract the properties of an image and enable the classification of images by predicting them based on the model it creates. Until today, many studies have been conducted in which diseases were diagnosed with the deep learning methods. In a previous study, deep learning methods have been used to detect and classify brain tumors from medical images [2]. In another study, recognizing the pathological characteristics of diabetic patients was provided with deep learning [3]. Besides, deep learning methods were used to diagnose thyroid nodules from ultrasound images [4].

## II. RELATED WORK

In medicine, diseases are diagnosed with the experience and knowledge of doctors. The use of an automatic diagnosis system can facilitate the work of doctors. Some studies [2, 3, 4, 5] use deep learning methods to diagnose eye diseases caused by diabetes to help diagnose and classify diseases. In the United States, the prevalence of diabetic retinopathy is approximately 28.5% among individuals that have diabetes, while this rate is 18% in India [5]. Most physicians refer their diabetic patients to the ophthalmologist at regular intervals for retinopathy or macular edema screening, depending on the severity of the disease. Automatic grading of diabetic retinopathy has the benefits of increasing efficiency, reproducibility, and the scope of screening programs. It can improve patient outcomes by reducing access barriers and providing early diagnosis and treatment. A type of Convolutional Neural Network (CNN) named Inception-v3 is generally used to aid image analysis and object detection. In

**İŞİL KARABEY AKSAKALLI**, is with Department of Computer Engineering University of Erzurum Technical University, Erzurum, Turkey, (e-mail: [isil.karabey@erzurum.edu.tr](mailto:isil.karabey@erzurum.edu.tr)).

 <https://orcid.org/0000-0002-4156-9098>

**SİBEL KAÇDIOĞLU**, is with Department of Computer Engineering University of Erzurum Technical University, Erzurum, Turkey, (e-mail: [sibel.kacdioglu@erzurum.edu.tr](mailto:sibel.kacdioglu@erzurum.edu.tr)).

 <https://orcid.org/0000-0003-0578-998X>

**YUSUF SINAN HANAY**, is with Department of Computer Engineering, University of Erzurum Technical University, Erzurum, Turkey, (e-mail: [sinan.hanay@erzurum.edu.tr](mailto:sinan.hanay@erzurum.edu.tr)).

 <https://orcid.org/0000-0002-3331-5936>

Manuscript received February 10, 2021; accepted April 19, 2021.  
DOI: [10.17694/bajece.878116](https://doi.org/10.17694/bajece.878116)

the EyePACS-1 data set, the sensitivity of the algorithm was 97.5%, and the specificity was 93.4%. In the Messidor-2 data set, the sensitivity was 96.1%, and the specificity was 93.9% [3]. In literature, CNN is also used for brain tumor detection. In [2], a deep learning-based brain tumor detection and classification system have been proposed using skull MR images. In the study, ELM-LRF (Local receptive field extreme learning machine) method was proposed for tumor classification. As a result of the experiments, the classification accuracy of MR images is 97.18%. The performance of the proposed method is better than recent studies conducted with commonly used methods such as CNN. In another study, a transfer learning method using the Inception-v3 model, which was previously adapted to medical image analysis, was proposed to classify nodules in the thyroid glands from ultrasound images [4]. By classifying 20 of the 21 FNA malignant glands as malignant, they obtained 95.2% sensitivity and 61.8% specificity values by classifying 21 of the 34 FNA benign glands as benign. Besides, in the external test set (100 gland appearance 50 benign, 50 malignant), 50 FNA classified 47 malignant glands as malignant and obtained 94% sensitivity, 50 FNA classified 28 benign glands and obtained 56% specific values [4]. Today, the fine needle aspiration ((FNA)) method is used when evaluating nodules. Computer-based nodule detection and classification can help doctors avoid unnecessary FNA.

There are many studies on the diagnosis and classification of kidney diseases by machine learning methods. In this study, a synthetic kidney function test (KFT) data set including age, gender, urea, creatinine, and glomerular filtration rate was created for the analysis of kidney disease. The study aims to compare the performance of the two methods under two headings as accuracy and working time by using the

the KNN classifier was found to be 89%, and the accuracy of the SVM classifier to 84% [7]. In a similar study, classification of kidney disease (stone or tumor) and segmentation was provided on ultrasound images [8]. Artificial neural networks are proposed for classification and multi-core k-means algorithm for segmentation. A median filter was used to remove noise in ultrasound images. GLCM (Level Co-occurrence Matrix) features were removed from each image after the noise was removed. As a result, it is seen that the system proposed as linear + quadratic-based segmentation has reached a maximum accuracy of 99.61% when compared to all other methods. Besides, the type of these stones is also important for treatment. To determine the type of stones, the type of kidney stone was classified from endoscopic video images with a deep learning network trained with digital photographs of five types of kidney stone components. This classification aims to automatically determine the laser energy settings manually adjusted according to the kidney stone component and size [9]. Deep convolutional neural network (CNN), as well as pre-trained ResNet-101, were used for classification. For recall values; UA (Uric acid) component 94.12%, COM (calcium oxalate monohydrate) component 90.48%, CHPD / brushite (calcium hydrogen phosphate dihydrate) component 85.71%, cystine (cystine) component 75%, MAPH / struvite (magnesium ammonium the phosphate hexahydrate) component is stated to be 71.42%. For precision values; UA (Uric acid) component 94.12%, COM (calcium oxalate monohydrate) component 95%, CHPD / brushite (calcium hydrogen phosphate dihydrate) component 75%, cystine (cystine) component 75%, MAPH / struvite (magnesium ammonium the phosphate hexahydrate) component is stated to be 71.43%. According to the researchers who conducted this study, it is the first study to

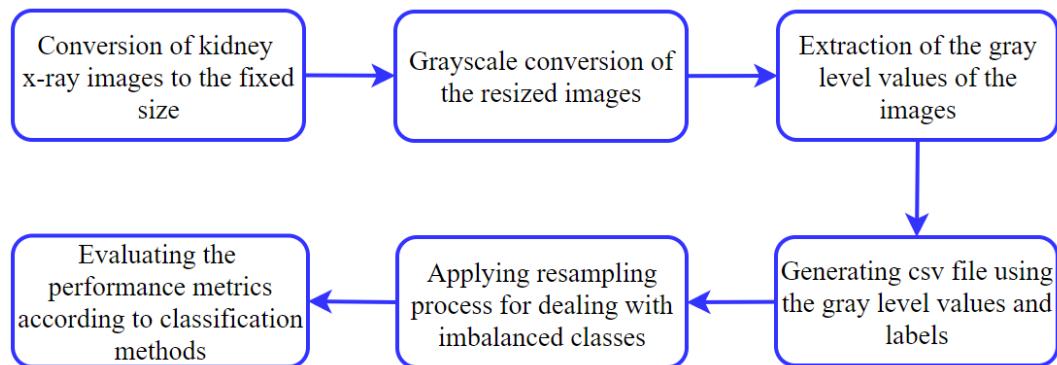


Fig.1. Block diagram of the proposed method

information of kidney patients and Support Vector Machine (SVM) and Artificial Neural Network (ANN) to predict four types of kidney disease [6]. Support Vector Machine (SVM) and Artificial Neural Network (ANN) methods were used. Classification accuracies were calculated as SVM 76.32%, ANN 87.70%. In another study, kidney stones are detected from low contrast ultrasound images. Median filter, Gaussian filter, and blunt masking are applied to improve the images. Subsequently, KNN and SVM classification techniques were used for the analysis of kidney stone images. The accuracy of

use a convolutional neural network to classify kidney stone type. In addition, the positions, shapes and sizes of kidney stones are different from each other. Therefore, kidney stone segmentation with machine learning is challenging. In the literature, preprocessing studies have been carried out to reduce this difficulty. In a study, a preprocess algorithm was developed for kidney stone detection and segmentation from CT images [10]. Three thresholding algorithms based on density, size, and location were applied to extract unrelated organ and bone structures from the images. CT images of 30

patients were studied. As a result, a 95.24% sensitivity value was obtained with the proposed algorithm [10]. In another study [11], the effects of morphological operations on kidney stone classification and analysis were investigated. The location and size of the kidney stone have been tried to be determined using GAC segmentation besides extraction and morphological operations. The proposed algorithms have been applied on several kidney images, and high efficiency has been achieved [11]. In one of the studies in the same direction, SVM was used for classification in automatic kidney stone detection. In the study, before classification, the image histogram equalization and embossing method, which evaluates color differences directionally, was tried. The proposed method was tested on 156 CT images with stones and healthy kidneys, achieving 98.71% accuracy [12]. In another study, a thresholding-based model has been developed with deep learning for the detection and scoring of kidney stones from abdominal non-contrast computed tomography (NCCT) images. The model is divided into four stages. Initially, 3D U-Nets were created for kidney and kidney sinus segmentation. Later, deep 3D dual-path networks were developed for hydronephrosis grading. Thresholding methods were used to identify and segment stones in the renal sinus area. Finally, the location of the stone was determined. As a result, the stone detection method reached 95.9% sensitivity and 98.7% positive predictive value (PPV) [13].

### III. OVERVIEW OF THE PROPOSED METHOD

Kidney x-ray images are used to detect whether a person has kidney stones or not. According to this information, a person is detected to be healthy or patient. This detection is generally decided by a specialist doctor. A sample of healthy or patient images is shown in Fig. 2. Although a specialist doctor can distinguish the images given in the figure, some images cannot be detected by the specialist, or the detection process by humans takes time. Therefore, an algorithmic detection system is needed for the classification of the x-ray images. In this study, a decision support mechanism that determines whether an individual is patient or healthy is proposed by applying various machine learning and deep learning methods to kidney x-ray images. The block diagram of the proposed mechanism is shown in Fig. 1. In the first step of Fig.1, each image was scaled to  $64 * 80$  dimensions because kidney x-ray images were obtained with different sizes. Then fixed-size images were converted into grayscale images to extract gray level values in the second step. In the third step, gray level binary values were extracted from these grayscale images, and these values were saved in a CSV file with their tags. Since the number of data with a healthy label in the data set is quite low compared to the patient label, resampling methods were applied to the dataset to deal with imbalanced classes. After all these processes were performed, various classification methods, including machine learning and deep learning have been applied to the balanced dataset, and the test dataset has been evaluated in terms of precision, recall, and F1 score.

#### A. Dataset

The dataset is prepared by using 221 kidney x-ray images obtained from the Urology Department of Ataturk University.

Before the classification process has been applied, these images are subjected to various preprocesses. In the first step, different-sized x-ray images are converted into  $64 \times 80$  fixed-size images. Then resized images are subjected to grayscale conversion processes in the second step. In the third step, the gray level values obtained from the image are extracted in the CSV file. These values consist of 5120 columns are labeled as patient or healthy according to the presence of kidney stones, catheters, or both found from the x-ray images. Images without any kidney stones or catheters are labeled as healthy, while images with stones or catheters are labeled as patients. This labeling process has been done by taking into account the opinions of the specialist doctors working in the Urology department. In the obtained dataset, 182 images have a patient label, while 39 images have a healthy label.

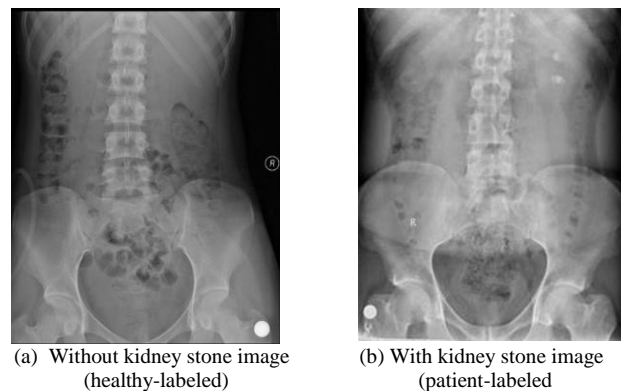


Fig. 2. Sample kidney x-ray images

#### B. Method

In this study, a decision support system based on machine learning and deep learning that detects whether a kidney x-ray image is patient or healthy is proposed. Machine learning (ML) is a branch of artificial intelligence, and it offers powerful classification techniques to make predictions on test data by training existing data and analyzing big data inaccessible to the human mind alone [14]. Machine learning methods are widely used mainly in data classification, pattern recognition, and prediction. Machine learning concepts are used for many applications such as data classification, email filtering, face detection, disease prediction, fraud detection, and traffic management. Deep learning (DL) is a type of machine learning method, and the learning process takes place on an artificial neural network model with more intermediate layers. Deep learning methods can classify large amounts of data with higher accuracy to provide analytical results based on the parameters and objectives of a particular framework [15]. Deep learning is mainly used in image segmentation, disease prediction, and recommendation systems such as convolutional neural networks, autoencoders, and restricted Boltzmann machines [16].

Within the scope of the study, various machine learning methods named Decision Tree (DT) [17], Random Forest (RF) [18], Support Vector Machine (SVM) [19], Multilayer Perceptron (MLP) [20], k Nearest Neighbor (kNN) [21] and Naive Bayes (BernoulliNB) [22] and deep learning

Convolutional Neural Network (CNN) which is a feed-forward neural network [23] have been applied. In the model training phase, the StratifiedKFold cross-validation method is applied to split the dataset. Also, the grid-search method is used to determine the best parameters belonging to the classification methods giving the highest accuracy rate

### 1) Resampling process

Since the data with the healthy label in the data used within the scope of the study are less than the data with the patient label, resampling is performed on the dataset. Among the resampling methods; undersampling, oversampling and SMOTETomek method, which is a combination of two methods, is applied in the scope of this study. Undersampling takes place by deleting a randomly selected section from the data belonging to the dominant class. In addition to bringing the data more stable, this method can shorten the running time of the classification method since it enables running with smaller data, especially when the data size is very large. On the other hand, as the information in the deleted data is lost, it can lead to underfitting problems. Oversampling can be defined as increasing the number of data belonging to this class by repeating a randomly selected part of the data belonging to the minority class. Since there is no loss of information in this method, it may be a superior method compared to undersampling. However, since some of the data are repeated precisely, it can lead to overfitting problems.

### 2) Classification

After resampling on the data set, train-test splitting is performed to be used as 80% training and 20% test data. The obtained train data is trained using StratifiedKFold cross-validation (kFold = 5). Then, test data that are not used in model training is evaluated in terms of Precision, Recall, and F1 score metrics. Performance percentages according to the resampling methods and classification algorithms used are given in Tables II and III.

TABLE I  
THE BEST PARAMETER VALUES FOR THE DATA SET ACCORDING TO MACHINE LEARNING METHODS

Algorithm	Best parameter values
Decision Tree (DT)	<i>criterion='gini', max_depth=3</i>
Random Forest (RF)	<i>criterion='entropy', max_depth= 4, max_features='log2', n_estimators= 200</i>
Support Vector Machine (SVM)	<i>C=0.1, gamma=1, kernel='rbf'</i>
Multilayer Perceptron (MLP)	<i>random_state=1, max_iter=300</i>
k Nearest Neighbor (kNN)	<i>n_neighbors= 16</i>
Naive Bayes (BernoulliNB)	<i>alpha= 10.0</i>

A general description of all classification methods applied within the scope of this study is given below:

**Decision Tree (DT):** Decision Tree is a supervised learning method generally used for classification and regression

analysis [24]. DT is expressed as a structural flow chart. Each internal node represents a test on an attribute, each branch describes a test result, and each leaf represents a class label [17, 25, 26].

**Random Forest (RF):** Random Forest, which consists of a combination of many decision trees, is used for more classification problems besides regression. A collection of tree-structured classifiers ( $h(x, \Theta_k), k = 1, \dots$ ) represents RF where  $\Theta_k$  are independent distributed random vectors and  $h_1(x), h_2(x)$ .. represents the training set obtained from random vector  $Y$ .  $X$  is a margin function that is calculated by  $mg(X, Y) = \text{av}_k I(h_k(X)=Y) - \max_{j \neq Y} \text{av}_k I(h_k(X)=j)$  where  $I(\cdot)$  is an indicator. The  $mg$  representing the margin gives the measure of the change in the average number of votes in  $X, Y$  with any other class. The larger  $mg$  value means, the more confidence in the classification [25].

**Support Vector Machine (SVM):** Support Vector Machine is a supervised machine learning method proposed by Vapnik [19]. SVM classifies the samples dividing the training dataset into distinct classes by using a hyperplane. When the dataset consists of two-dimensional data, a linear classifier is used with a linear hyperline [19, 27]. In this study, we use the linear classifier to distinguish the healthy and patient-labeled data. In the mathematical expression,  $x$  is a vector point, and  $w$  is a weight. This classifier aims to find the optimal plane where the distance between the two classes is the greatest to keep the margin value at the highest level. This case is called the maximum margin linear classifier, and the calculation of the margin is expressed in Fig. 3 and equation 1.

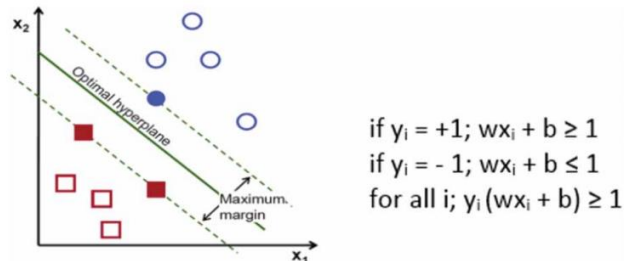


Fig. 3. Linear classification using SVM

$$\text{margin} \equiv \arg \min_{x \in D} d(x) = \arg \min_{x \in D} \frac{|\mathbf{x} \cdot \mathbf{w} + b|}{\sqrt{\sum_{i=1}^d w_i^2}} \quad (1)$$

**Multilayer Perceptron (MLP):** Multilayer Perceptron is a popular feed-forward neural network due to its fast operation, easy applicability, and small dataset requirements [26]. In this network structure, units consist of an input layer, one or more hidden layers, and an output layer [20]. The input layer takes an activation vector externally and transmits it to the units in the first hidden layer via weighted links. After each layer calculates its activation, it transfers the activation to the neurons in the successive layers, as shown in Fig. 4.

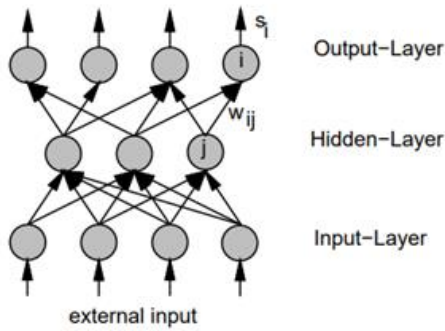


Fig. 4. Multilayer Perceptron with one hidden layer

Each neuron  $i$  in the network is a simple processing unit that calculates the activation  $s_i$  based on incoming excitation called the  $net_i$ .  $net_i$  is calculated as seen in equation 2:

$$net_i = \sum_{j \in pred(i)} s_j w_{ij} - \theta_i \quad (2)$$

In this equation,  $pred(i)$  indicates the set of predecessors of unit  $i$ , while  $w_{ij}$  represents the weight of the connection from unit  $j$  to unit  $i$ .  $\theta_i$  is the bias value of the unit  $i$ . The activation of the unit  $i$  is calculated by passing the net input through a nonlinear activation function. Usually the sigmoid logistic function is calculated as follows:

$$s_i = f_{log}(net_i) = \frac{1}{1 + e^{-net_i}} \quad (3)$$

Having an easily computable derivative of this function makes the method advantageous. The derivation is shown below:

$$\frac{\partial s_i}{\partial net_i} = f'_{log}(net_i) = s_i * (1 - s_i) \quad (4)$$

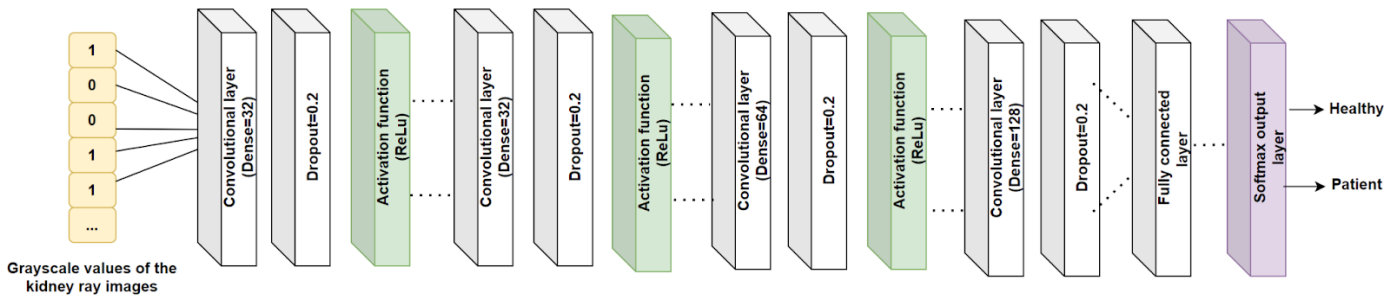


Fig.5. Designed CNN model for classification of kidney x-ray images

*k Nearest Neighbor (kNN)*: kNN is a supervised learning method that estimates test data based on the samples closest to  $k$  values given in the feature space [27, 28]. After training all existing samples, the method classifies new samples according to the similarity measure. As a result of the experiments performed with the Grid Search method, the optimal  $k$  value is found as 16.

*Naive Bayes (BernoulliNB)*: Naive Bayes method calculates the probability of finding the correct tag of a data from a test dataset by multiplying the probabilities of all factors affecting that result [29]. In the equation below;  $C$  is the class label,  $F$  values are the input data:

$$P(C | F_1, F_2, \dots, F_n) = \frac{P(C) P(\bar{F}_1, \bar{F}_2, \dots, F_n | C)}{P(F_1, F_2, \dots, F_n)} \quad (5)$$

with the operation, there is a probability value that the input data whose class is unknown belongs to the  $C$  class. As a result of probability calculations, the number of class labels is determined as the class of the test data [29].

*Convolutional Neural Network (CNN)*: CNN is one of the most famous successful methods that has been widely used in image processing in recent years. This method is based on artificial neural networks, whose network structure contains more intermediate layers, neurons with learnable weight, and bias. CNN input data, which consists of any image or digital data, differs according to the problem in terms of the dropout value, the number of layers, the activation function used, and the number of neurons [30]. Within the scope of this study, the CNN model is created using Keras library and Python programming language. The best result has been tried to be achieved by changing the learning rate, optimization algorithm, number of hidden layers, number of epochs, weight starting values, and activation functions as hyperparameters. Fig. 5 shows the developed CNN model to classify whether the person is healthy or patient.

#### IV. EXPERIMENTAL RESULTS

After the resampling process is completed, 80% of the dataset (182 image values) is trained, and precision, recall, and F1 score performance metrics are evaluated with the StratifiedKFold cross-validation method. The classification performance of the methods depends on the number of correctly detected classes (TP-Correct Positive), the number of healthy people identified as patients (FP-False Positive), and the number of patients identified as healthy (FN-False Negative). Using these values, the Precision and Recall values

are calculated. F1 score value gives the harmonic average of the precision and recall values. Therefore, a high F1 score value is an essential criterion for a suitable decision support mechanism. General formulas of metrics used in the study are given in equations (6) - (8).

$$\text{Precision} = TP / (TP + FP) \quad (6)$$

$$\text{Recall} = TP / (TP + FN) \quad (7)$$

TABLE II CROSS VALIDATION SCORE FOR TRAINING MODEL (S: SMOTE, U: RANDOMUNDERSAMPLER, ST: SMOTETOMEK, S+U: COMBINIG SMOTE – RANDOMUNDERSAMPLER)

	Cross_val_score											
	Precision (%)				Recall (%)				F1 Score (%)			
	S	U	ST	S+U	S	U	ST	S+U	S	U	ST	S+U
DT	85.8	88	83.8	90.2	85.8	88	83.8	90.2	85.5	70	77.5	77.2
RF	85.5	85.5	85.9	85.8	85.5	85.5	85.5	85.9	91	81.6	87	87.5
SVM	85.8	85.8	85.8	85.8	85.8	85.8	85.8	85.8	92.4	92.4	92.4	92.4
MLP	84.7	88.3	87.2	87.6	84.7	88.3	87.2	87.6	90.3	77.6	85	82.7
kNN	86.2	84.6	80.7	87.4	86.2	84.6	80.7	87.4	87	88.6	32.8	66.6
Bayes	85.8	85.8	79.3	85.8	85.8	85.8	79.3	85.8	92.4	92.4	27.2	92.4
CNN	77.9	77.9	77.9	77.9	100	100	100	100	87.4	87.4	87.4	87.4
<b>AVG</b>	<b>84.5</b>	<b>85.1</b>	<b>82.9</b>	<b>85.6</b>	<b>87.7</b>	<b>88.3</b>	<b>86.1</b>	<b>88.9</b>	<b>89.4</b>	<b>84.3</b>	<b>69.9</b>	<b>83.7</b>

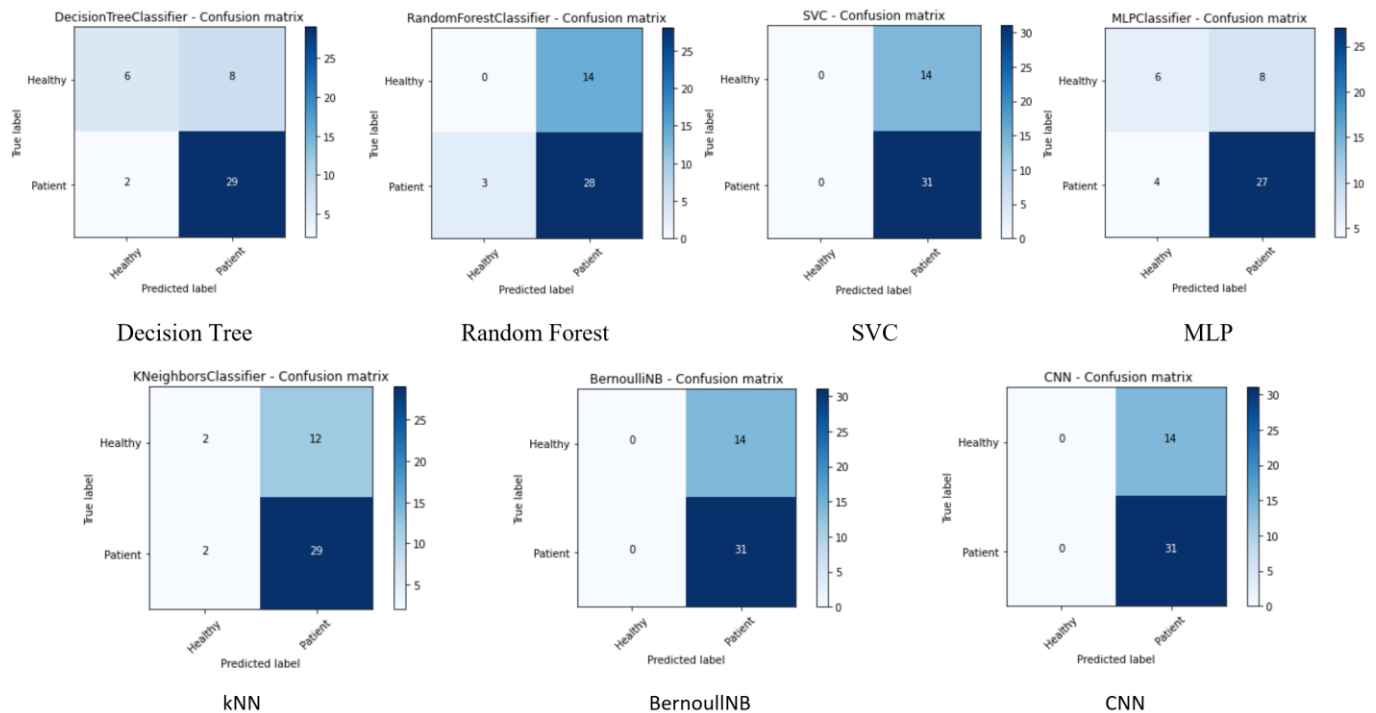


Fig. 6. Confusion matrices obtained from combining SMOTE+RandomUnderSampler method according to classification methods

$$F1\_score = 2 * (Precision * Recall) / (Precision + Recall) \quad (8)$$

Table II shows the performances of the cross-validation method applied to the training model according to algorithms and sampling methods. While the average highest values in terms of precision (85.6%) and recall (88.9%) metrics are obtained with the combining SMOTE+ RandomUnderSampler (S + U) method, when evaluated in terms of algorithms, the average highest value for F1 Score (89.4%) is obtained using SMOTE sampling method. According to cross-validation scores of the applied algorithms, MLP and CNN achieve the highest values in terms of the evaluation metrics. MLP has an 87.2% precision rate and 92.4% F1 score rate, while CNN has a 100% recall rate.

Table III shows the performance values by applying the cross\_val\_predict method to 45 images that have not been used before on the trained model. In terms of precision, the highest rate (78.4%) is obtained with the Decision Tree Classifier (DT) using the S + U method. When looking at the results of the prediction for the test data (cross\_val\_predict), 100% Recall value is obtained with the CNN method in all samples, while the Precision and F1 score values as a result of ST and S + U sampling are quite low compared to other methods. Furthermore, it is seen that the Decision Tree Classifier (DT) has the highest F1 score by 85.3.1% rate using S+U.

Also, this method gives the highest precision rate (78.4%)

TABLE III CROSS VALIDATION PREDICT FOR TEST MODEL (S: SMOTE, U: RANDOMUNDERSAMPLER, ST: SMOTETOMEK, S+U: COMBINIG SMOTE – RANDOMUNDERSAMPLER)

	Cross_val_predict											
	Precision (%)				Recall (%)				F1 Score (%)			
	S	U	ST	S+U	S	U	ST	S+U	S	U	ST	S+U
DT	72.7	80	69.6	<b>78.4</b>	77.4	77.4	51.6	93.5	75	78.7	59.3	<b>85.3</b>
RF	70.7	68.6	63.9	66.7	93.5	77.4	74.2	90.3	80.6	72.7	68.7	76.7
SVM	68.9	68.9	70.3	68.9	<b>100</b>	<b>100</b>	83.9	<b>100</b>	81.6	81.6	76.5	81.6
MLP	71.9	72.7	76.5	77.1	74.2	77.4	83.9	87.1	73	75	80	81.8
kNN	72.5	68.2	33.3	70.7	93.5	96.8	6.5	93.5	81.7	80	1.08	80.6
Bayes	68.9	68.9	70.4	68.9	<b>100</b>	<b>100</b>	61.3	<b>100</b>	81.6	81.6	65.5	81.6
CNN	68.9	68.9	68.9	68.9	<b>100</b>	<b>100</b>	<b>100</b>	<b>100</b>	81.6	81.6	81.6	81.6
<b>AVG</b>	<b>70.6</b>	<b>70.9</b>	<b>64.7</b>	<b>71.3</b>	<b>91.2</b>	<b>89.7</b>	<b>65.9</b>	<b>94.1</b>	<b>79.3</b>	<b>78.7</b>	<b>61.8</b>	<b>81.3</b>

among the other methods. According to both Table II and Table III, it is examined that all of the sampling methods give the same results in the CNN method.

When the methods are evaluated in terms of sampling, it is seen that the S + U combined sampling strategy gives the highest average recall (94.1%) value, Precision (71.3 %), and F1 score (81.3 %) values. Fig.6 shows complexity matrices obtained as a result of the classification of test data using S+U oversampling. When the confusion matrices seen in Fig. 3 are examined, it is seen that the BernoulliNB and CNN models classify the patient data 100%. Therefore the recall value obtained from these models is 100%. However, these methods couldn't correctly classify the healthy labeled data.

## V. CONCLUSION

In this study, kidney x-ray images obtained from Atatürk University Research Hospital are used to classify patients and healthy individuals implementing machine learning and deep learning approaches. By using these methods, a decision support mechanism is proposed in a shorter time that enables the diagnosis of images that the specialist doctor has difficulty in diagnosing. Firstly, images are converted to the gray level values after they are scaled to fixed sizes. Then, a data set is created by obtaining gray-level numerical values from the images. Since this data set has imbalance classes, various oversampling and undersampling methods are used. In this way, the performance metrics of the methods increase significantly. Accurate detection of healthy individuals is as important as the detection of patient individuals in the detection of kidney diseases. In this respect, achieving high performance in the F1 score is one of the most important criteria. According to the experiments, DT has the highest F1 score rate with a success rate of 85.3% using the S+U sampling method.

## ACKNOWLEDGMENT

The authors thank the Department of Urology at Ataturk University Education and Research Hospital to share kidney x-ray images anonymously,

## REFERENCES

- [1] D. Aune, Y. Mahamat-Saleh, T. Norat, and E. Riboli, "Body fatness, diabetes, physical activity and risk of kidney stones: a systematic review and meta-analysis of cohort studies," *European journal of epidemiology*, vol. 33, no. 11, pp. 1033-1047, 2018.
- [2] A. Ari and D. Hanbay, "Deep learning based brain tumor classification and detection system," *Turkish Journal of Electrical Engineering & Computer Sciences*, vol. 26, no. 5, pp. 2275-2286, 2018.
- [3] V. Gulshan et al., "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *Jama*, vol. 316, no. 22, pp. 2402-2410, 2016.
- [4] J. Song et al., "Ultrasound image analysis using deep learning algorithm for the diagnosis of thyroid nodules," *Medicine*, vol. 98, no. 15, 2019.
- [5] R. Raman et al., "Prevalence of diabetic retinopathy in India: Sankara Nethralaya diabetic retinopathy epidemiology and molecular genetics study report 2," *Ophthalmology*, vol. 116, no. 2, pp. 311-318, 2009.
- [6] S. Vijayarani, S. Dhayanand, and M. Phil, "Kidney disease prediction using SVM and ANN algorithms," *International Journal of Computing and Business Research (IJCBR)*, vol. 6, no. 2, pp. 1-12, 2015.
- [7] J. Verma, M. Nath, P. Tripathi, and K. Saini, "Analysis and identification of kidney stone using K th nearest neighbour (KNN) and support vector machine (SVM) classification techniques," *Pattern Recognition and Image Analysis*, vol. 27, no. 3, pp. 574-580, 2017.
- [8] A. Nithya, A. Appathurai, N. Venkatadri, D. Ramji, and C. A. Palagan, "Kidney disease detection and segmentation using artificial neural network and multi-kernel k-means clustering for ultrasound images," *Measurement*, vol. 149, p. 106952, 2020.
- [9] K. M. Black, H. Law, A. Aldouhki, J. Deng, and K. R. Ghani, "Deep learning computer vision algorithm for detecting kidney stone composition," *BJU international*, 2020.
- [10] N. Thein, H. A. Nugroho, T. B. Adji, and K. Hamamoto, "An image preprocessing method for kidney stone segmentation in CT scan images," in *2018 International Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM)*, 2018: IEEE, pp. 147-150.
- [11] M. Shahina and H. S. Mahesh, "Renal Stone Detection and Analysis by Contour Based Algorithm," in *2019 International Conference on Recent Advances in Energy-efficient Computing and Communication (ICRAECC)*, 2019: IEEE, pp. 1-5.
- [12] A. Soni and A. Rai, "Kidney Stone Recognition and Extraction using Directional Emboss & SVM from Computed Tomography Images," in *2020 Third International Conference on Multimedia Processing, Communication & Information Technology (MPCIT)*, 2020: IEEE, pp. 57-62.
- [13] Y. Cui et al., "Automatic Detection and Scoring of Kidney Stones on Noncontrast CT Images Using STONE Nephrolithometry: Combined

- Deep Learning and Thresholding Methods," *Molecular Imaging and Biology*, pp. 1-10, 2020.
- [14] L. E. McCoubrey, M. Elbadawi, M. Orlu, S. Gaisford, and A. W. Basit, "Harnessing machine learning for development of microbiome therapeutics," *Gut Microbes*, vol. 13, no. 1, pp. 1-20, 2021.
- [15] B. Shah, A. Alsadoon, P. Prasad, G. Al-Naymat, and A. Beg, "DPV: a taxonomy for utilizing deep learning as a prediction technique for various types of cancers detection," *Multimedia Tools and Applications*, pp. 1-23, 2021.
- [16] A. K. Sahoo, C. Pradhan, and H. Das, "Performance evaluation of different machine learning methods and deep-learning based convolutional neural network for health decision making," in *Nature inspired computing for data science*: Springer, 2020, pp. 201-212.
- [17] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81-106, 1986.
- [18] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5-32, 2001.
- [19] V. N. Vapnik, "An overview of statistical learning theory," *IEEE transactions on neural networks*, vol. 10, no. 5, pp. 988-999, 1999.
- [20] M. W. Gardner and S. Dorling, "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences," *Atmospheric environment*, vol. 32, no. 14-15, pp. 2627-2636, 1998.
- [21] D. W. Aha, D. Kibler, and M. K. Albert, "Instance-based learning algorithms," *Machine learning*, vol. 6, no. 1, pp. 37-66, 1991.
- [22] K. P. Murphy, "Naive bayes classifiers," *University of British Columbia*, vol. 18, no. 60, 2006.
- [23] K. Fukushima and S. Miyake, "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition," in *Competition and cooperation in neural nets*: Springer, 1982, pp. 267-285.
- [24] K. C. Kandpal, S. Kumar, G. S. Venkat, R. Meena, P. K. Pal, and A. Kumar, "Onsite age discrimination of an endangered medicinal and aromatic plant species *Valeriana jatamansi* using field hyperspectral remote sensing and machine learning techniques," *International Journal of Remote Sensing*, vol. 42, no. 10, pp. 3777-3796, 2021.
- [25] M. A. Friedl and C. E. Brodley, "Decision tree classification of land cover from remotely sensed data," *Remote sensing of environment*, vol. 61, no. 3, pp. 399-409, 1997.
- [26] S. Naaz, "Detection of Phishing in Internet of Things Using Machine Learning Approach," *International Journal of Digital Crime and Forensics (IJDCF)*, vol. 13, no. 2, pp. 1-15, 2021.
- [27] M. Riedmiller, "Advanced supervised learning in multi-layer perceptrons—from backpropagation to adaptive learning algorithms," *Computer Standards & Interfaces*, vol. 16, no. 3, pp. 265-278, 1994.
- [28] E.-H. S. Han, G. Karypis, and V. Kumar, "Text categorization using weight adjusted k-nearest neighbor classification," in *Pacific-asia conference on knowledge discovery and data mining*, 2001: Springer, pp. 53-65.
- [29] G. Yang et al., "Tree Species Classification by Employing Multiple Features Acquired from Integrated Sensors," *Journal of Sensors*, vol. 2019, 2019.
- [30] I. Karabey and L. Bayındır, "An evaluation of fingerprint-based indoor localization techniques," in *2015 23rd Signal Processing and Communications Applications Conference (SIU)*, 2015: IEEE, pp. 2254-2257.
- [31] İ. K. AKSAKALLI and L. BAYINDIR, "Derin Öğrenme Kullanılarak Parmak izi Tabanlı İç Ortam Konumlandırma," *Erzincan Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, vol. 13, no. 2, pp. 483-501.

## BIOGRAPHIES



**ISIL KARABEY AKSAKALLI** graduated from Gazi University in 2013 and started to work as a research assistant at Atatürk University in 2014. After receiving her MSc degree from Atatürk University in 2015, she was appointed as a research assistant to Erzurum Technical University in the

Department of Computer Engineering. She started her PhD. in 2016 at Hacettepe University and still continues to this program. Her research topics include microservice architectures, optimization methods, distributed systems, machine learning and deep learning techniques.



**SIBEL KACDIOGLU** obtained a bachelor degree in Computer Engineering from Anadolu University in 2016. She received her M.Sc. in 2020 in Computer Engineering Department at the Ataturk University, Erzurum. Currently she pursues her Ph.D. in Computer Engineer at Computer Engineering Department of Ataturk University, Erzurum. Along with her Ph.D. she works as a research assistant in Erzurum Technical University. Her areas of interest are Machine Learning, Computer Vision and Image Processing.



**Y. SINAN HANAY** received Ph.D. in Electrical and Computer Engineering from University of Massachusetts (UMass), Amherst in 2011. He joined Osaka University in 2011 as a post-doctoral researcher. From 2012 to 2016, he worked at Center for Information and Neural Networks (CiNet) of NICT Japan. During that time, he also held visiting researcher and lecturer positions at Osaka University. He is a co-author of a paper which received the best paper award in IEEE International Conference on High Performance Switching and Routing 2011. He is currently an assistant professor at Erzurum Technical University in Turkey.