

DIAGNOSIS OF THE DISEASES USING RESAMPLING METHODS WITH MACHINE LEARNING ALGORITHMS

Ahmet Çelik

Received on December 14, 2022

Presented by K. Atanassov, Member of BAS, on May 30, 2023

Abstract

The rapid diagnosis of diseases is very important for the early start of the treatment process. Pneumonia is a disease that affects the lungs and can cause death in advanced cases. Pneumonia is still today diagnosed by doctors while examining chest X-ray images. As diagnosis of the diseases using machine learning algorithms will be useful. In addition, high success of rate will be obtained while classifying balanced dataset by using machine learning algorithms. Resampling methods are used to balance the dataset by using under-sampling or over-sampling methods. In literature, there is no study comparing under-sampling and over-sampling methods. In the study, an open source dataset was used which included two classes published through the Kaggle data store. The data set includes 1341 healthy and 3875 pneumonia chest X-ray images. Two different resampling methods named Random under-sampling (RU) and ADASYN (Adaptive Synthetic) over-sampling were used while balancing healthy and pneumonia images. After this operation obtained data were used for training machine learning algorithms. In the study, first and second level attributes of the X-ray images in the dataset were used. Logistic Regression (LR) and Support Vector Machine (SVM) algorithms were used for classification of dataset. According to the results obtained, 93.109% accuracy rate of classification success was achieved of X-ray image dataset which balanced by ADASYN over-sampling method with SVM algorithms.

Key words: machine learning, pneumonia, medical imaging techniques, medical decision, chest X-ray images, classification

Introduction. Pneumonia is a life-threatening infectious respiratory disease usually caused by bacteria, viruses or fungi. In addition, it was observed that COVID-19 causes pneumonia in the lungs in the advanced stage during the pandemic. The intelligent systems have become widespread to diagnose diseases by using computer vision methods. Machine learning or deep learning algorithms are used for classification of diseases data in intelligent systems. These algorithms are sub-fields of artificial intelligence. It is possible to design fast and accurate decision-making systems by processing many data in the field of health. Many studies have been conducted to diagnose disease using patient and symptoms datasets. But there is no study which balanced disease dataset by under-sampling and over-sampling methods and compared resampling methods.

YAHYAOUÏ and YUMUSAK [1] determined the prediction success of certain chest diseases such as pneumonia and asthma using Deep Neural Network (DNN) and k-Nearest Neighbours (kNN) methods. According to the results obtained 95% by using kNN and 94.3% by using DNN correct result detection was achieved.

JAIN et al. [2] presented Convolutional Neural Networks for pneumonia detection using X-ray images. In the study several Convolutional Neural Networks were trained to classify pneumonia and non-pneumonia data. In addition 6 models were also mentioned. According to the results obtained, the highest success rate of 98% recall metric value was obtained.

MALL et al. [3] applied different machine learning approaches to classify bone X-ray images of MURA (Musculoskeletal Radiographs) dataset into fractures and no fractures categories. They used LBF SVM (Radial Basis Function support vector machine), linear SVM, Logistic Regression (LR) and Decision Tree (DT) algorithms for abnormality detection using data obtained from automatic attribute method. According to the results obtained, the highest success rate of 60.86% was obtained by the Logistic Regression algorithm.

MOURA et al. [4] investigated classification models to distinguish between COVID-19-based and typical chest X-ray images of pneumonia using hand-made radiographic features of COVID-19. Chest X-ray images used for testing were rescaled to 256×256 pixels. In the study SVM, RF, AdaBoost (AB) and LR algorithms were used for classification. According to the results obtained, the highest success rate of 98% accuracy metric value was obtained by using AdaBoost algorithm.

CHOWDARY [5] used features extracted by an automatic feature extraction method called VGG16 to diagnose pneumonia from chest X-ray images. In the study chest X-ray images were converted to gray scale format and rescaled to 224×224 pixels. RF, LR and SVM machine learning algorithms were used for classification. According to the results obtained, the highest success rate of 95% accuracy metric value was obtained.

HAÏBO et al. [6] applied ADASYN, SMOTE and DT resampling methods on six datasets from the UCI Machine Learning Repository and compared perfor-

mance of methods. According to the results obtained the best performance in all metric by using ADASYN method in the study.

In this study, a dataset containing 1341 healthy chest X-ray images and 3875 pneumonia chest X-ray images was used from the open source Kaggle repository. Detection of pneumonia success was measured with LR and SVM machine learning algorithms using balanced methods on the dataset. The imbalance of the dataset directly affects the classification success while using machine learning algorithms. In the study, the under-sampling and over-sampling methods were used to balance the training data classes of X-ray images dataset. In the first method, Random under-sampling was used to reduce and equalize the number of samples of Healthy and Pneumonia classes. The performances of success were measured after trained 1341 Healthy and 1341 Pneumonia chest X-ray images by using machine learning algorithms. In the second method, ADASYN over-sampling was used to increase and equalize the number of samples Healthy and Pneumonia classes. The performances of success were measured after trained 3875 Healthy and 3875 Pneumonia chest X-ray images by using machine learning algorithms.

Material and method. All operations were applied by using three steps in the study. The diagram of the system is shown in Fig. 1. In the first step, Random under-sampling and ADASYN over-sampling methods were used to balance the class sets of the X-ray images taken from the data set. In the second step, image preprocessing methods were used to determine the boundaries of the lung region on chest X-ray images and total of 8 attributes were extracted from the chest X-ray images manually. In the third step, pneumonia and healthy chest X-ray image classification was performed with LR and SVM machine learning algorithms using attributes obtained manually from image processing methods and the success of performance was determined.

The flow diagram of the system shown in Fig. 1a used the Random under-sampling method. In this method, 1341 Pneumonia chest X-ray images were obtained by under-sampling randomly from 3875 pneumonia chest X-ray images to balance training class data. Thus, 1341 healthy chest and 1341 pneumonia chest X-ray images were used for training data.

The flow diagram of the system shown in Fig. 1b used the ADASYN over-sampling method. In this method, 3875 healthy chest X-ray images were obtained by over-sampling from the 1341 healthy chest X-ray images to balance training class data. Thus, 3875 healthy chest and 3875 pneumonia chest X-ray images were used for training data.

Preprocessing of images. In this study, first, pixel intensities in the image were determined using Histogram Equalization. In the second step, Mask R-CNN was used on the X-ray image with pixel intensity to perform segmentation of the lung region. In the last preprocessing step, the lung region was converted to black and white colours using the Otsu thresholding algorithm. After this stage, first and second-order textural features were determined.

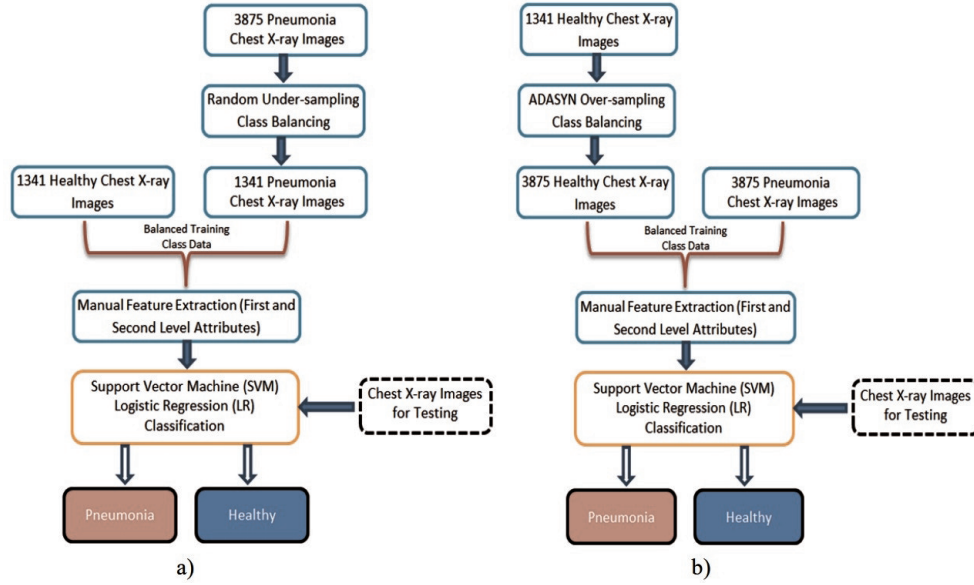


Fig. 1. The diagram of resampling methods: a) Random under-sampling and classification using SVM and LR algorithms; b) ADASYN over-sampling and classification using SVM and LR algorithms

First-level features. The ratio of white pixels to black pixels, mean and variance were used as first-order features. After preprocessing the image, the features were extracted on a pixel-based level. The obtained feature values were obtained only from the limited lung region in the X-ray image.

Second-level features. Gray Level Co-occurrence Matrix (GLCM) was calculated using the values of $d = 1$ and $\theta = 0, 45, 90, 135$ [7]. The ASM (Angular Second-Moment), Correlation, Contrast, Homogeneity and Difference were calculated from the Chest X-ray image. The Scikit-Image Python library was used to calculate the five features [8].

SVM (Support Vector Machine) machine learning algorithm. Support Vector Machine (SVM) is a machine learning algorithm developed by Cortes and Vapnik to solve regression and classification problems [5]. SVM is an algorithm based on optimization mathematics used in classification problems. The goal of SVM is to find an optimal linear hyper plane that best classifies data examples for linear classification. The margins between hyper planes passing through the support vectors of both classes are maximized, and there are no data examples in this margin [9]. In this way, test data gains the ability to make the best prediction. Although SVM is an algorithm used in binary separation problems, multiclass problems are addressed by using one versus all methods [10].

Logistic regression machine learning algorithm. Logistic Regression (LR) is a statistical model developed by Tolles and Meurer for regression and

classification tasks [5]. Unlike linear regression, LR performs classification tasks. In a binary classification problem, σ represents the label that feature vector $y = \{0, 1\}$ can take. LR finds the probability of which class the x feature vector belongs to [11]. The sigmoid function is calculated to find the probability by using Eq. (1) [12].

$$(1) \quad \sigma(t) = \frac{1}{1 + e^{-t}}.$$

Resampling methods for imbalanced classes. Balancing the classes in the dataset is achieved by reducing or increasing the number of class examples. Class sample reduction (under-sampling) is performed by equalizing the number of other classes in the dataset with the minimum class sample count. Class sample increase (over-sampling) is done by making the number of other classes in the dataset equal to the highest number of the class.

As the number of sample classes increases, more synthetic examples are generated to increase the overall learning performance. However, in some machine learning algorithms, increasing the class instance can improve learning performance. In methods that increase the number of class samples, accurate results may not always be obtained when the data between classes is not linear. In addition, over fitting occurs due to the repeated addition of data that makes class decisions more stringent.

ADASYN (Adaptive Synthetic) and SMOTE (Synthetic Minority Over-sampling Technique) techniques are commonly used to solve classification imbalance problems in datasets and increase the number of samples [13]. Similarly, both methods use the k-nearest neighbour algorithm. In addition, Random under-sampling is a commonly used method to address class imbalance and sample reduction.

Random under-sampling method. In supervised learning, multi-class imbalanced data classification is one of the most challenging research topics in machine learning. Traditional machine learning algorithms often show a bias towards the majority class examples, while ignoring the minority class examples. As a result, ignoring minority class examples affects the accuracy of classifiers [14]. In the Random under-sampling method, random removal is performed by the majority class examples. This method has the ability to contribute positively to classification performance [15]. A classifier trained with imbalanced data can produce biased results in favour of the majority class. One way to solve this problem is to use the Random sub-sampling method to obtain a more balanced class distribution [16]. The Random under-sampling method forms the basis of classification with machine learning methods [17].

ADASYN over-sampling method. The situations of class imbalance in datasets are quite common in real-world problems. Class imbalance occurs when the number of training classes in the dataset used in the classification process

differ significantly from each other [18]. The over-sampling methods can also be used to solve classification imbalance. ADASYN method is one of the common methods used to solve class imbalance in datasets [19].

In the ADASYN method, as the bias is reduced and the decision boundary is shifted towards harder examples to learn, the learning performance increases. The number of synthetic samples to be generated is determined using the beta coefficient. A complete balance between the classes is achieved by using $\beta=1$.

Using density distribution, a different number of synthetic samples are generated for each sample [20]. In the first step of the ADASYN algorithm, the degree of class imbalance is calculated by using Eq. (2).

$$(2) \quad d = \frac{m_s}{m_l},$$

where d , is the degree of class imbalance, m_s is the number of samples in the minority class and m_l is the number of samples in the majority class. In the second step, the number of synthetic samples to be produced for the minority class is calculated by using Eq. (3).

$$(3) \quad G = (m_s - m_l) \times \beta.$$

Here G , is the number of samples needed, $\beta \in [0, 1]$, is a coefficient indicating the equilibrium level. In the third step, the ratio is calculated by using Eq. (4).

$$(4) \quad r_i = \Delta_i / K \quad i = 1, \dots, m_s.$$

Here K , is the number of nearest neighbours for i -th minority class example. Δ_i is the number of samples belonging to the majority class within the nearest neighbours. In the fourth step, the r_i ratio is normalized. This normalization process is calculated by using Eq. (5).

$$(5) \quad \hat{r}_i = r_i / \sum_{i=1}^{m_s} r_i.$$

In the fifth step, the number of synthetic samples to be produced for each sample belonging to the minority class is calculated by using Eq. (6).

$$(6) \quad g_i = \hat{r}_i \times G,$$

where g_i , is the number of synthetic samples to be produced for the i -th sample. In the sixth step, a loop is created for a minority class instance and g_i synthetic instance is produced in this loop. This loop starts from 1 and continues until g_i . The loop is calculated by using Eq. (7).

$$(7) \quad s_i = x_i + (x_{zi} - x_i) \times \lambda,$$

where x_i , is the i -th minority class instance, x_{zi} , is a minority class sample selected randomly from the k nearest neighbours of x_i , s_i , is the generated synthetic sample, λ , is a random number between 0 and 1.

Results and discussion. In this study, X-ray images obtained from Kaggle storage area were balanced by using Random under-sampling and ADASYN over-sampling methods. In the study, the healthy and pneumonia chest classes from dataset were balanced for getting higher accuracy results. The first and second-level features were extracted after preprocessing methods were applied to the dataset images. As first-level features Mean, Variance and the ratio of the number of white pixels to the number of black pixels were used. As second-level attributes Contrast, correlation, homogeneity, ASM, Dissimilarity values were used. By using these features, classification of healthy and pneumonia chest X-ray images was performed with SVM and LR algorithms.

In the study, first and second-level features of the datasets obtained by Random under-sampling and ADASYN over-sampling methods on machine learning algorithms were compared based on success. Confusion Matrix, metric graphics and Receiver Operating Characteristic (ROC) curves of the obtained results were created. The Random under-sampling technique was used to reduce the number of chest X-ray samples with pneumonia from 3875 to 1341 in the dataset. The ADASYN over-sampling technique was used to increase the number of chest X-ray samples with healthy from 1341 to 3875 in the dataset.

Confusion Matrix values of SVM and LR machine learning algorithms on Random under-sampling and ADASYN over-sampling dataset are shown in Fig. 2. Both first and second-level features were used in the testing phase. Using Random under-sampling with the SVM algorithm, 231 of 234 chest X-ray images of healthy and 338 of 390 chest X-ray images of pneumonia were detected correctly and shown in Fig. 2a. With the LR algorithm, 226 of 234 chest X-ray images of healthy and 312 of 390 chest X-ray images of pneumonia were detected correctly shown in Fig. 2b. According to the results the SVM algorithm classified both healthy and pneumonia X-ray images with higher success.

Furthermore, using ADASYN over-sampling with the SVM algorithm, 224 of 234 chest X-ray images of healthy and 357 of 390 chest X-ray images of pneumonia were detected correctly and shown in Fig. 2c. With the LR algorithm, 230 of 234 chest X-ray images of healthy and 335 of 390 chest X-ray images of pneumonia were detected correctly and shown in Fig. 2d. According to the results the SVM algorithm classified healthy X-ray images more successfully, and LR algorithm classified X-Ray images with pneumonia more successfully.

The classification success graphic of the SVM and LR machine learning algorithms using both the first and second-level features of the dataset obtained using the Random under-sampling and the ADASYN over-sampling method are shown in Fig. 3. The Accuracy, Precision, Recall, F1-Score and AUC Score metrics for the SVM and LR algorithms are shown in the graphic separately. Using Random

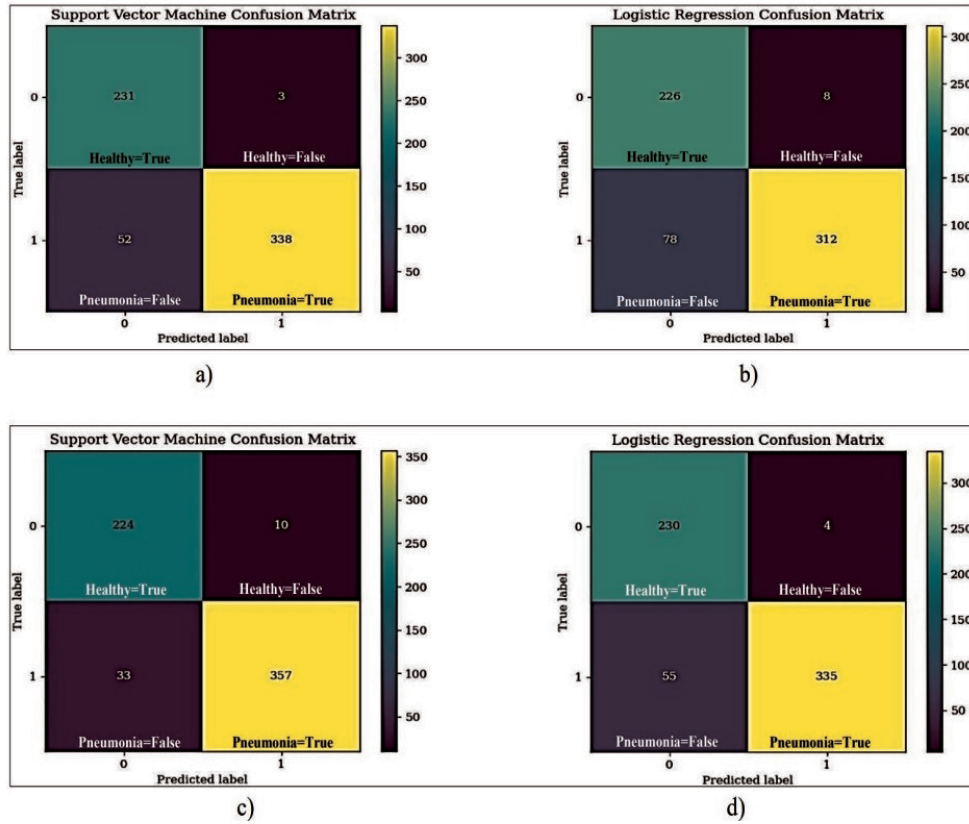


Fig. 2. Confusion Matrix success values with the first and second-level attributes using ML algorithms and resampling methods: a) SVM algorithm success using the Random under-sampling method; b) LR algorithm success using the Random under-sampling method; c) SVM algorithm success using the ADASYN over-sampling method; d) LR algorithm success using the ADASYN over-sampling method

under-sampling method, the Accuracy metric value was found to be 91.15% with the SVM algorithm and 86.217% lower with the LR algorithm; and the AUC Score metric value was found to be 97.737% with the SVM algorithm and 95.805% lower with the LR algorithm. Furthermore, using ADASYN over-sampling method, the Accuracy metric value was found to be 93.109% with the SVM algorithm and 90.544% with the LR algorithm; and the AUC Score metric value was 98.388% with the SVM algorithm and 97.534% with the LR algorithm. Similar success rates were obtained for other metric values. According to the results ADASYN over-sampling method outperformed Random under-sampling method on all metrics while using two algorithms.

The ROC curve graphic is shown in Fig. 4. Figure 4a shows the ROC curve of both the first and second-level feature success rates of the dataset obtained using the ADASYN over-sampling method with SVM and LR machine learning

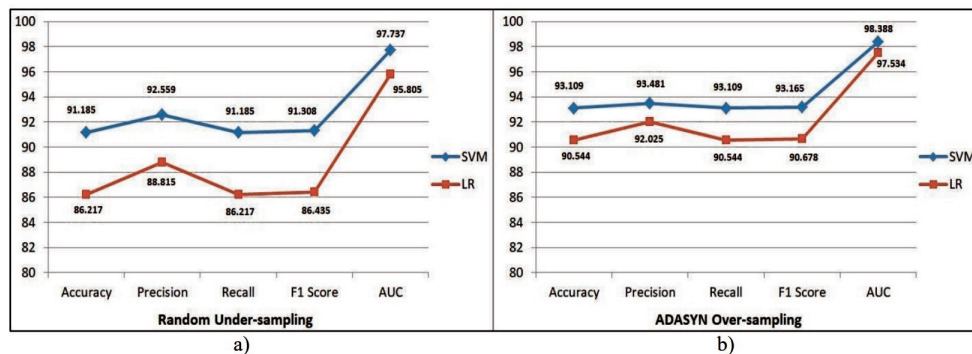


Fig. 3. Metric graphics and success rate of resampling methods: a) Metric graphic and success rate using Random under-sampling method; b) Metric graphic and success rate ADASYN over-sampling method

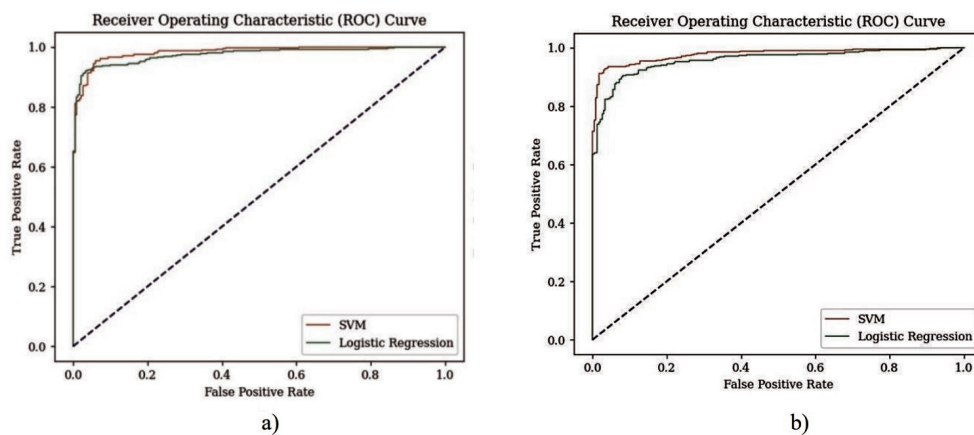


Fig. 4. Success curves of SVM and LR algorithms with resampling methods: a) The ROC curve graphic of ADASYN over-sampling method; b) The ROC curve graphic of Random under-sampling method

algorithms. Figure 4b presents the ROC curve of both the first and second-level feature success rates of the dataset obtained using the Random under-sampling method with SVM and LR machine learning algorithms. In both graphs, the ROC curve of the SVM algorithm is more stable than the ROC curve of the LR algorithm.

In this study, Random under-sampling and ADASYN over-sampling class balancing methods for training data were used. The success rates of SVM and LR algorithms in pneumonia disease diagnosis were determined using manually generated first and second-level attributes of chest X-ray images. In the first stage, using Random under-sampling method, 1341 normal and 1341 pneumonia chest X-Ray images were used for training data. Using ADASYN over-sampling method, 3875 normal and 3875 pneumonia chest X-ray images were used for training data.

In the second stage, Histogram equalization, Otsu threshold and mask R-CNN preprocessing methods were performed after using resampling methods and first and second-level attributes were extracted from dataset. In the last stage, SVM and LR machine learning algorithms were used for healthy and pneumonia chest classification.

Conclusion. According to the results obtained SVM algorithm performed better than LR algorithm while using two resampling methods. By using Random under-sampling method 97.737% AUC success rate was performed with SVM algorithm. In addition, by using ADASYN over-sampling method 98.388% AUC success rate was performed with SVM algorithm. In addition, it was also observed that SVM algorithm performed better classification than LR algorithm using all performance metrics.

In conclusion, both resampling methods were analyzed and tested by SVM and LR machine learning algorithm. As a result of tests a higher success rate was obtained by using SVM algorithm. However, ADASYN over-sampling method showed better performance than Random under-sampling method. It was observed that the classification performance increased after using the ADASYN over-sampling method. The reason for this is that the more data the algorithms are trained with, the more the success rate to increase. In academic literature reviews, most of the articles preferred to increase the data in order to overcome the data imbalance problem. However, there is no study comparing data reduction and increase methods. For this reason, this study will be an important and original source of literature that can be used in future studies.

REFERENCES

- [1] YAHYAOU A., N. YUMUSAK (2021) Deep and machine learning towards pneumonia and asthma detection, Proceedings of 2021 International Conference on Innovation and Intelligence for Informatics, Computing and Technologies, 2021, Zallaq, Bahrain, 494–497.
- [2] JAIN R., P. NAGRATH, G. KATARIA (2020) Pneumonia detection in chest X-Ray images using convolutional neural networks and transfer learning, Measurement, **165**, 108046.
- [3] MALL P. K., P. S. SINGH, D. YADAV (2019) GLCM based feature extraction and medical X-RAY image classification using machine learning techniques, Proceedings of 2019 IEEE Conference on Information and Communication Technology, Allahabad, India, 1–6.
- [4] MOURA L. V., C. M. DARTORA, C. M. OLIVEIRA, R. C. BARROS, A. M. M. SILVA (2020) A novel approach to differentiate COVID-19 pneumonia in chest X-Ray, Proceedings of IEEE 20th International Conference on Bioinformatics and Bioengineering, Cincinnati, OH, USA, 446–451.

- [5] CHOWDARY G. J. (2021) Impact of machine learning models in pneumonia diagnosis with features extracted from chest X-Rays using VGG16, *Turk. J. Com. Math. Ed.*, **12**, 1521–1530.
- [6] HAIBO H., B. YANG, G. EDUARDO, L. SHUTAO (2008) ADASYN: Adaptive synthetic sampling approach for imbalanced learning, *Proceedings of the International Joint Conference on Neural Networks*, Hong Kong, 1322–1328.
- [7] HARALICK R., K. SHANMUGAM, I. DINSTEN (1973) Textural features for image classification, *IEEE Trans. Syst. Man Cyb.*, **3**, 610–621.
- [8] VAN DER WALT S., J. L. SCHÖNBERGER, J. N. IGLESIAS, F. BOULOGNE, J. D. WARNER et al. (2022) Scikit-image: image processing in Python, 2014, <https://peerj.com/articles/453/> (Accessed: June 2022).
- [9] CERVANTES J., F. G. LAMONT, L. R. MAZAHUA, A. L. CHAU (2020) A comprehensive survey on support vector machine classification: Applications, challenges and trends, *Neurocomputing*, **408**, 189–215.
- [10] FARIS H., M. HABIB, M. FARIS, M. ALOMARI, A. ALOMARI (2020) Medical speciality classification system based on binary particle swarms and ensemble of one vs. rest support vector machines, *J. Bio. Inf.*, **109**, 103525.
- [11] KSIAZEK W., M. GANDOR, P. PLAWIAK (2021) Comparison of various approaches to combine logistic regression with genetic algorithms in survival prediction of hepatocellular carcinoma, *Comp. Bio. and Med.*, **134**, 104431.
- [12] ZOU X., Y. HU, Z. TIAN, K. SHEN (2019) Logistic regression model optimization and case analysis, *Proceedings of the IEEE 7th International Conference on Computer Science and Network Technology (ICCSNT)*, Dalian, China, 135–139.
- [13] LETTERI I., A. CECCO, A. DYOUB, G. D. PENNA (2022) A novel resampling technique for imbalanced dataset optimization, 2012, <https://arxiv.org/pdf/2012.15231.pdf> (Accessed: May 2022).
- [14] ARAFAT M. Y., S. HOQUE, S. XU, D. M. FARID (2019) An under-sampling method with support vectors in multi-class imbalanced data classification, *Proceedings of the 13th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, Island of Ulkulhas, Maldives, 1–6.
- [15] BRANCO P., L. TORGO, R. P. RIBEIRO (2022) A survey of predictive modelling under imbalanced distributions, <https://arxiv.org/abs/1505.01658> (Accessed: April 2022).
- [16] PRUSA J., T. M. KHOSHGOFTAAR, D. J. DITTMAN, A. NAPOLITANO (2015) Using random under-sampling to alleviate class imbalance on tweet sentiment data, *Proceedings of the 2015 IEEE International Conference on Information Reuse and Integration*, California, USA, 197–202.
- [17] WANG H., Y. LIU, Y. CHANGCHUN, X. REN, C. JIN et al. (2021) Fast 3D time-domain airborne EM forward modeling using random under-sampling, *J. Appl. Geo.*, **191**, 104357.
- [18] PRISTYANTO Y., A. F. NUGRAHA, A. DAHLAN, L. A. WIRASAKTI, A. A. ZEIN et al. (2022) Multiclass imbalanced handling using ADASYN oversampling and stacking algorithm, *Proceedings of the 16th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, Seoul, Korea, 1–5.
- [19] ZHANG J., S. LI, Y. DING, J. TANG, F. GUO (2020) An two-layer predictive model of ensemble classifier chain for detecting antimicrobial peptides, 2020 *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Seoul, Korea, 56–61.

- [²⁰] HASMITA S., F. NHITA, D. SAEPUDIN, A. ADITSANIA (2019) Chili Commodity Price Forecasting in Bandung Regency using the Adaptive Synthetic Sampling (ADASYN) and K-Nearest Neighbor (KNN) Algorithms, Proceedings of the 2019 International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 434-438.

*Computer Technologies Department
Tavşanlı Vocational School
Kütahya Dumlupınar University
Kütahya, Turkey
e-mail: ahmet.celik@dpu.edu.tr*