IJAMEC
e-ISSN: 3023-4409

INTERNATIONAL JOURNAL OF APPLIED METHODS IN ELECTRONICS AND COMPUTERS

www.ijamec.org

International Open Access

Volume 12
Issue 01

March, 2024

Research Article

Boosting the classification success in imbalanced data of bee larva cells

Serkan ÖZGÜN [a], Mehmet Akif ŞAHMAN [b]

[a] *Selcuk University, Department of Computer Engineering, Campus, Selcuklu, Konya, Turkey*
[b] *Selcuk University, Department of Electrical and Electronics Engineering, Campus, Selcuklu, Konya, Turkey*

ARTICLE INFO

Article history:
Received 28 November 2023
Accepted 06 January 2024

Keywords:
Imbalanced dataset,
Oversampling method
SMOTE
Beekeeping
Detection of larva cells

ABSTRACT

Selecting the appropriate honey harvesting method is crucial for sustainable beekeeping and optimal honey production. The use of primitive harvesting methods can lead to the death of bees and a decrease in honey yield. This study aims to address the issue of detecting and classifying young larvae on honeycombs. However, the area where young larvae are found is limited compared to other areas. In this study, the dataset obtained from honeycombs was imbalanced, which has used the Synthetic Minority Oversampling TEchnique (SMOTE) algorithm to balance it. The SMOTE algorithm is a synthetic data generation method. The balanced dataset was then used for classification processes with k-Nearest Neighbors algorithm (k-NN), Decision Trees, and Support Vector Machines. The evaluation of the classification results included the F1-Score, G-Mean, and AUC metrics. The results showed that the classification of the dataset balanced with synthetic data was more successful.

1. Introduction

An imbalanced dataset occurs when a class is dominant over other classes. [1]. Machine learning and data mining concepts have raised the problem of the accuracy of the evaluated results. This situation paved the way for studies to stabilize imbalanced datasets. Although there were studies on these issues before, the foundation for the work regarding imbalanced datasets was laid in early 2000. In 2003 the issue of an imbalanced dataset became a subject of study on its own [2]. An imbalance in the number of data between classes problem is a common problem in datasets and classification and also is often encountered in real-world problems. [3]. Confidentiality in the data collection process, costly data collection, and the low number of samples are the main reasons for the imbalanced dataset problem [4]. The issue of imbalanced datasets arises in various real-world scenarios, including credit card fraud detection [5], identification of defective products in production facilities [6], inaccurate medical diagnoses [7], detection of unauthorized access in computer networks [8], identification of oil spills in the ocean [9], text classification problems [10], and software error prediction [11].

The imbalance between the data of the classes causes the majority class to dominate the results. [12]. While classifiers are proposed, they are generally studied on balanced data sets. Therefore, in imbalanced datasets, the classifiers' success rate has decreased. Balancing the dataset is important in order to increase classification success before evaluating imbalanced datasets. [13]. The operation for balancing an imbalanced dataset can be categorized into 3 steps. These approaches are data level, algorithm level, and hybrid level (data + algorithmic level). Balancing the dataset at the data level can be based on two pillars. The first is to reduce the number of majority class instances that are known as under-sampling, and the second, increase minority class samples which is known as over-sampling [14].

One of the biggest problems in honey production is losing young larvae during harvesting. If all honey in the comb is desired to be collected during the honey filtering phase, then unfortunately the young baby bees found on the comb also perish. In this case, to achieve the maximum level of honey output, beekeepers often give up on the offspring, which leads to bee loss and unhomogenized

* Corresponding author. E-mail address: asahman@selcuk.edu.tr
DOI: 10.58190/ijamec.2024.78

honey. This situation causes the bee colony to weaken in the long run. On the other hand, not thoroughly filtering the honeycombs causes a decrease in honey production. In this study, successful classification and detection of larva cells are implemented through proper classification studies.

Figure 2 shows that the larvae are found in a limited area compared to other areas, and in some areas, there are no larva cells. This situation causes an imbalanced dataset. To address this imbalanced dataset issue, the synthetic minority oversampling technique (SMOTE) is used.

The study compares the results of the original dataset with the balanced data using SMOTE. Classifiers such as Decision Tree (DT), k-Nearest Neighbors (k-NN), and Support Vector Machine (SVM) are employed. The used evaluation metrics are F1-Score, G-Mean, and AUC. The literature is given in section 2, the materials and methods used in this study are presented in section 3, the results obtained and discussion are analyzed in section 4, and the conclusions in this paper are summarized in section 5.

2. Literature Review

To this day many solutions have been produced to the problem of classifying imbalanced datasets. Hart [15], proposed ignoring samples that are worthless and far from the boundary line, using the CNN (Condensed Nearest Neighbor) method [16]. Tomek [17], proposed the Tomek-Link method. Tomek - Link is where majority classes are categorized under the same classes as their closest neighbors [18]. Kubat and Matwin [19], studied with the One-Sided Selection method which was somewhat based on the Tomek-Link method [18]. Wilson [20], improved the Edited Nearest Neighbor algorithm. In the ENN method, the three closest neighbors of a sample are selected. If two of the selected neighbors are from the opposite class, this instance is deleted. Later, a method based on this method was developed in which only the majority of class samples were removed from the dataset [21]. Mani and Zhang [22], proposed the Nearmiss method which was based on the k-NN method.

Chawla et al. [23], proposed one of the most popular oversampling methods called the SMOTE. SMOTE produces synthetic data similar to minority class samples.

After the SMOTE algorithm was proposed, shortcomings of the technique were found, and different methods have been proposed to resolve it. Han et al. [24], came up with the Borderline- SMOTE method. Borderline-SMOTE is based on the SMOTE technique. However, this method does not consider all minority class instances, it only considers borderline examples of the minority class [25]. They divided this method into two sub-methods, Borderline-SMOTE1 and Borderline-SMOTE2. In the Borderline-SMOTE2 method, the majority class samples were also considered along with the minority class samples [3]. Bunkhumpornpat et al. [25], later proposed the Safe-Level-SMOTE method. With this method, synthetic data is produced only in the safe zone. He et al. [26] proposed the ADASYN method. In this method, minority class samples were given a weight value according to their learning disability. More data is generated from the samples that have more weight than this weight value [27].

3. Materials and Methods

Langstroth and Dadant are the most used hive types in the world. Both types have many similarities however one differs in dimension. In Turkey, the Longstroth hive type is mostly used. In this study, the hive that are used and accepted as standard is the Langstroth type. Langstroth hive-type is suitable for beekeeping activities and is suitable for the climatic conditions of our country. The Langstroth hive type consists of 5 parts: bottom board, hive body, honey super, inner cover, and cover.

3.1. Materials

In this study, data obtained from 19 different honeycombs with Langstroth standard are used. Since images are taken from both sides of the honeycombs, a total of 38 honeycomb images are obtained.

Images are obtained from the BASLER acA2500-14uc scanning camera. Obtained images are 2590 x 1940 pixels. However, only the parts containing the honeycomb region are taken, images with a size of 1162 x 574 pixels, seen in Figure 1, are obtained. The resulting images are divided into 5 x 5 pixel pieces. 26.448 images are obtained from each hive.

Cells on a honeycomb can be labelled as such, cells with pollen, closed cells with larvae, closed cells with honey, open cells with some honey, hollow cells, and open cells with larva. The images in this study can be labelled as areas with larvae on them, and areas with no larva. Images containing closed larva cells are labelled as “1”, other images are labelled as “0”. In the determination of these areas, closed cells containing larvae are displayed on a 5x5 pixel image. If the area covered is more than other areas, it is labelled as “1”.

A total of 1.005.024 thumbnail images are achieved from 38 honeycomb images. 816.877 of the images are from the class labelled as "0", and 188.144 of them are cells containing larvae labelled “1”. As seen in Figure 2, 5x5 pixel split images of the honeycomb and the R-G-B values are entered and recorded into the columns. Thus, for each image, there are 75 columns with values between 0 and 255 in one column. 1.005.024 rows of data with 76 columns, including the label value, are obtained (Figure 3).

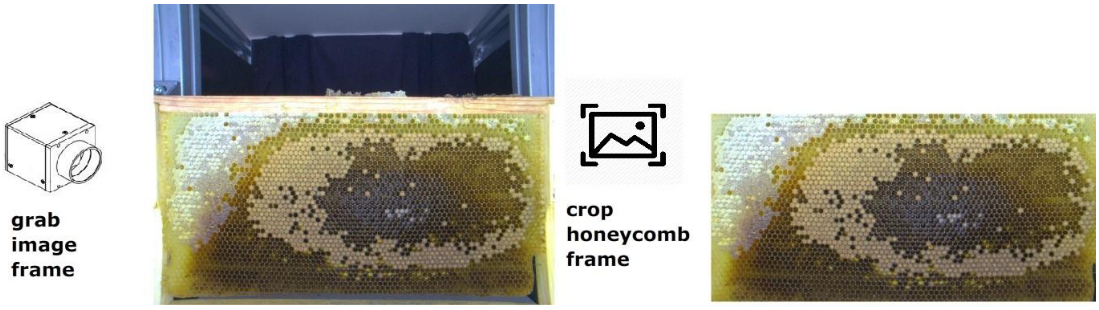


Figure 1. First stage of honeycomb image

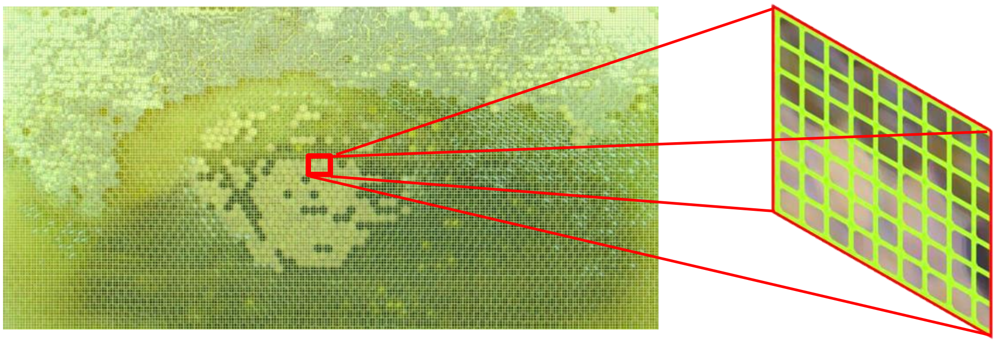
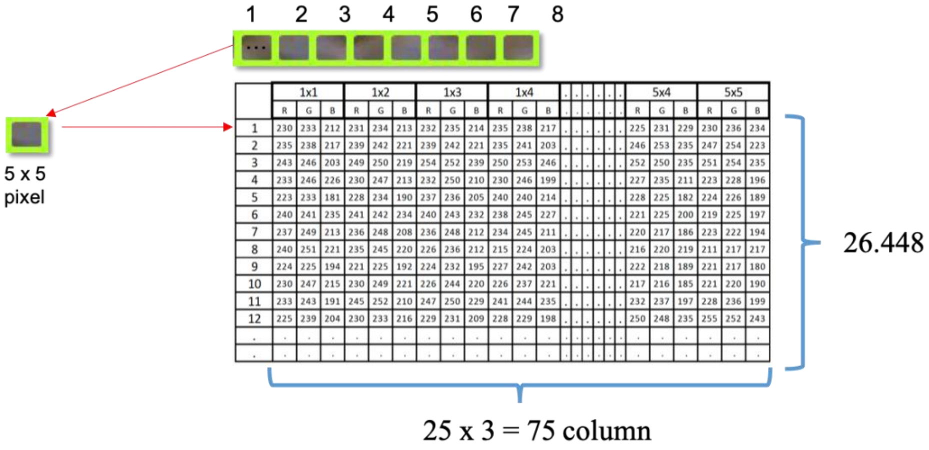


Figure 2. 5x5 pixel fragmented honeycomb image



	1x1			1x2			1x3			1x4			.	.	.	.	.	.	5x4			5x5		
	R	G	B	R	G	B	R	G	B	R	G	B	.	.	.	.	.	.	R	G	B	R	G	B
1	230	233	212	231	234	213	232	235	214	235	238	217	.	.	.	.	.	.	225	231	229	230	236	234
2	235	238	217	239	242	221	239	242	221	235	241	203	.	.	.	.	.	.	246	253	235	247	254	223
3	243	246	203	249	250	219	254	252	239	250	253	246	.	.	.	.	.	.	252	250	235	251	254	235
4	233	246	226	230	247	213	232	250	210	230	246	199	.	.	.	.	.	.	227	235	211	223	228	196
5	223	233	181	228	234	190	237	236	205	240	240	214	.	.	.	.	.	.	228	225	182	224	226	189
6	240	241	235	241	242	234	240	243	232	238	245	227	.	.	.	.	.	.	221	225	200	219	225	197
7	237	249	213	236	248	208	236	248	212	234	245	211	.	.	.	.	.	.	220	217	186	223	222	194
8	240	251	221	235	245	220	226	236	212	215	224	203	.	.	.	.	.	.	216	220	219	211	217	217
9	224	225	194	221	225	192	224	232	195	227	242	203	.	.	.	.	.	.	222	218	189	221	217	180
10	230	247	215	230	249	221	226	244	220	226	237	221	.	.	.	.	.	.	217	216	185	221	220	190
11	233	243	191	245	252	210	247	250	229	241	244	235	.	.	.	.	.	.	232	237	197	228	236	199
12	225	239	204	230	233	216	229	231	209	228	229	198	.	.	.	.	.	.	250	248	235	255	252	243
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.

Figure 3. Converting the R-G-B values of the 1 honeycomb image to the dataset.

3.2. Method

3.2.1. Synthetic Minority Oversampling Technique (SMOTE)

One of the most known and used data sampling algorithms was developed by Chawla et al. known as SMOTE [23]. The algorithm uses the minority class samples synthetically to find solutions to imbalanced datasets[28]. Synthetic data generation of the SMOTE approach is shown in Figure 4. In the SMOTE algorithm, the k nearest neighbours (x_j) of a minority sample (x_i) are selected. The distance difference between the sample and the selected neighbours is calculated. A randomly chosen (α) value between 0 and 1 is multiplied by the obtained difference value. As can be seen in Equation 1, the calculated value is added to the sample itself, and a new synthetic sample (x_{new}) is generated[13].

$$x_{new} = x_i + (x_i - x_j) * \alpha \tag{1}$$

3.2.2. Classification Methods

3.2.2.1. k Nearest Neighbour (k-NN)

The k-NN algorithm is the most popular among machine learning classification methods.

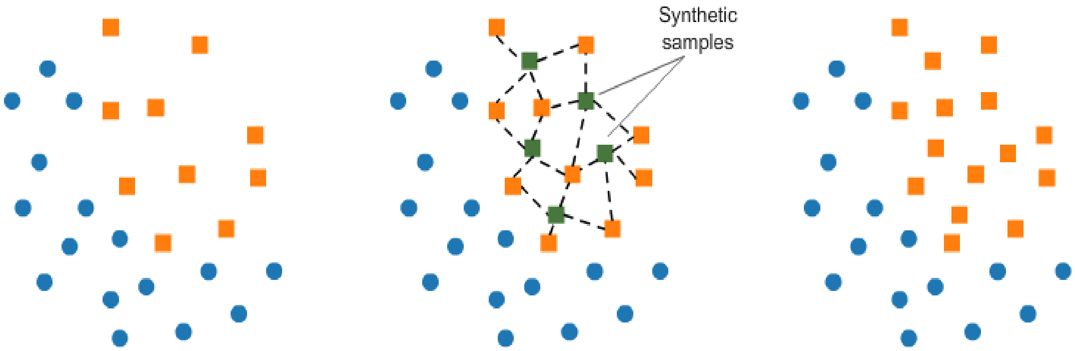


Figure 4. SMOTE algorithm

It is one of the simple classification methods. The k-NN algorithm is sample-based and is used to solve classification problems. For the classification process, k pieces of data belonging to the sample are selected according to the distance criteria. As distance measures, Minkowski, Euclid, Chebyshev, and cosine equations are used, Euclidean distance is often used in the literature. Whichever class the selected k data is more included, the sample to be classified belongs to that class [29].

3.2.2.2. Decision Tree (DT)

One of the most used methods in classification problems is the DT algorithm. In the DT method, each attribute is represented by a node. Branches and leaves are parts of a tree structure. Root at the top, leaves at the bottom, and branches in between [30]). In DT algorithms, feature classes are combined into a single class. The main purpose is to divide it into two until one class remains. The sample to be classified will progress until it reaches a leaf between branches. At the point it reaches the leaf, the class it represents accepts the class of the instance [31].

3.2.2.3. Support Vector Machines (SVM)

SVM is a statistical learning-based algorithm that performs controlled classification. The SVM method aims to find the most appropriate boundary between different samples by dividing the data into two or more classes with linear mechanisms in 2-dimensional space, planar in 3-dimensional space, and hyper-plane mechanisms in multidimensional space[32]. Hyper-plane detection is easy in linear classes. However, hyper-plane detection is difficult in nonlinear classes. The non-linearly separated sample space is moved to the upper space where the samples can be linearly separated and the hyper-plane is found[29].

3.3. Evaluation Metrics

The selection of suitable evaluation metrics is crucial for accurate evaluations. The main criterion used when evaluating the classification results is accuracy. However, using the accuracy criteria alone may be insufficient to be an indicator of success in datasets. Therefore, the imbalanced dataset is mainly measured with; F1-Score, G-Mean, and AUC (Area). To determine the metrics, first of all, the Confusion Matrix seen in Table 1 needs to be obtained.

Table 1: Confusion Matrix

		Predictive Values	
		Positive	Negative
Actual Values	Positive	TP	FN
	Negative	FP	TN

True Positives (TP) : Correctly predicted positive samples

False Positives (FP) : Incorrectly predicted negative samples

True Negatives (TN) : Correctly predicted negative samples

False Negatives (FN) : Incorrectly predicted positive samples

To evaluate the metrics, we must first calculate the following equations using the confusion matrix [13].

$$Accuracy = (TP + TN)/(TP + FP + TN + FN) \quad (2)$$

$$Precision = TP/(TP + FP) \quad (3)$$

$$Recall = TP/\ (TP + FN) \quad (4)$$

$$Specificity = TN/\ (TN + FP) \quad (5)$$

The F1-Measure is a more reliable measure because it takes the harmonic average of the sensitivity and precision measures instead of taking a simple average. The equality of the F1-Score is specified in Equation 6 [13].

$$F1 - Score = 2 \times (Precision \times Recall)/(Precision + Recall) \quad (6)$$

The G-Mean measure shows the balance between the two classes (Equation 7). While F1-Score evaluates the success of the minority class belonging to the imbalanced dataset, G-Mean evaluates for all classes [3].

$$G - Mean = \sqrt{Recall * Specificity} \quad (7)$$

The AUC-ROC curve is one of the most important metrics for measuring classification success, especially in imbalanced datasets. The ROC curve is a curve with a true positive rate (Sensitivity) on its vertical axis and a false positive rate (Specificity) on its horizontal axis (Figure 5). A near-perfect result should have a ROC curve from (0,0) to (0,1) that is almost vertical and then from (0,1) to (1,1) almost horizontally[33].

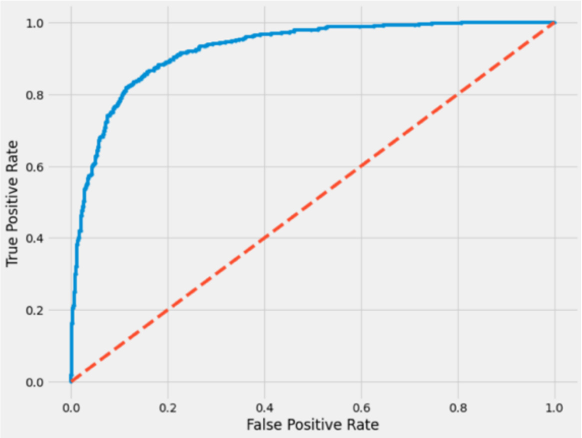


Figure 5: AUC-ROC Curve

AUC corresponds to the area under the ROC curve. The size of the AUC area is directly proportional to the classification success. The ideal result for AUC is 1.

4. Results and Discussion

To obtain the results in this study, the honeycomb images were fragmented and labelled. A dataset with 1.005.024 rows and 76 columns is used. Dataset with k-fold cross validation is divided into 5 subsets by random selection of rows. The indices of randomly selected subsets are saved for later use of the subset to which the samples belonged. Then, classification operations are applied to our original data. For classification, k-NN, DT, and SVM algorithms are used. For evaluation, F1-Score, G-Mean, and AUC metrics are used in the operations.

In the second stage, synthetic data were generated for the minority class using the SMOTE algorithm. Thus, 628.730 synthetic data belonging to the minority class are generated, thus the minority class and the majority class are balanced. Synthetic data obtained are again divided into 5 subsets and merged with previous subsets. Therefore, divided into 5 subsets 1.633.754 data are obtained and this data is reclassified. The obtained data are shown in Table 2.

When the classification results are evaluated, it is seen that the classification success increases after the dataset is balanced with synthetic data generation for all classification algorithms and all metrics. When the classification results made with the k-NN algorithm are evaluated according to the F1-Score metric, the classification success, which is 92.36%, is 97.56% and a significant increase is observed.

Table 2: Classification results obtained with real data and synthetic data

	Classification /Metric	**k-NN**	**DT**	**SVM**
Original Data	F1-Score, %	**92.36**	87.71	16.9
	G-Mean, %	**96.97**	92.26	33.94
	AUC, %	**96.97**	92.39	42.16
SMOTE	F1-Score, %	**97.56**	96.55	38.69
	G-Mean, %	**97.47**	96.54	43.91
	AUC, %	**97.50**	96.54	53.06

The success of the k-NN algorithm in the F1-Score is also valid for the DT algorithm. The success, which is 87.71%, has become 96.55% after synthetic data is generated. However, the difference in the percentage of success in the G-Mean and AUC metrics is greater than the k-NN algorithm. In the DT algorithm, serious success is observed in all metrics in general.

It is seen that the classification success is low in the results obtained in the SVM algorithm. However, after applying the SMOTE algorithm, it is seen that the classification success has increased significantly. Again, in the performance measurement make with the F1-Score, the success percentage, which is 16.9%, has become 42.16%.

5. Conclusions

In this study, the problem of imbalanced datasets is addressed. To balance imbalanced datasets, SMOTE, which is one of the well-known and most used synthetic data generation methods in the literature, was used. Classification results were evaluated with classification methods that are frequently used in machine learning and data mining.

Detection of the areas with larvae on the honeycombs is an important point in terms of increasing honey yield. After balancing the imbalanced dataset obtained from the existing honeycombs with the SMOTE algorithm, the obtained dataset was subjected to classification processes and the classification results were evaluated. In the results obtained, it was seen that the classification success was better in the dataset balanced with synthetic data generation.

According to the results obtained, in the k-NN classification results; there was an increase of 5.2% according to the F1-Score metric, 0.5% according to the G-Mean metric, and 0.53% according to the AUC metric. In the results obtained with the DT classification; An increase in classification success was observed, such as 8.84% in the F1-Score metric, 4.28% in the G-Mean metric, and 4.15% in the AUC metric. In the SVM classification, an increase of 21.79% according to the F1-Score metric, 9.97% according to the G-Mean metric, and 10.9% according to the AUC metric was observed.

In future studies, the existing dataset can be oversampled with different synthetic data generation methods used in the literature and comparisons can be made with different classification methods.

Acknowledgment

This study has been presented in 2nd International Symposium on Implementations of Digital Industry and Management of Digital Transformation (ISIDIMDT'21), 10-11 November 2021, Konya/Turkey.

References

[1] N. V. Chawla, N. Japkowicz, and A. Kotcz, "Special issue on learning from imbalanced data sets," *ACM SIGKDD explorations newsletter,* vol. 6, no. 1, pp. 1-6, 2004.

[2] A. Fernández, S. Garcia, F. Herrera, and N. V. Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *Journal of artificial intelligence research,* vol. 61, pp. 863-905, 2018.

[3] E. Kaya, S. Korkmaz, M. A. Sahman, and A. C. Cinar, "DEBOHID: A differential evolution based oversampling approach for highly imbalanced datasets," *Expert Systems with Applications,* vol. 169, p. 114482, 2021.

[4] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of imbalanced data: A review," *International journal of pattern recognition and artificial intelligence,* vol. 23, no. 04, pp. 687-719, 2009.

[5] M. Zareapoor and J. Yang, "A novel strategy for mining highly imbalanced data in credit card transactions," *Intelligent Automation & Soft Computing,* pp. 1-7, 2017.

[6] S. Cateni, V. Colla, and M. Vannucci, "A method for resampling imbalanced datasets in binary classification tasks for real-world problems," *Neurocomputing,* vol. 135, pp. 32-41, 2014.

[7] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," *Neural networks,* vol. 21, no. 2-3, pp. 427-436, 2008.

[8] D. A. Cieslak, N. V. Chawla, and A. Striegel, "Combating imbalance in network intrusion datasets," in *GrC*, 2006: Citeseer, pp. 732-737.

[9] M. Kubat, R. C. Holte, and S. Matwin, "Machine learning for the detection of oil spills in satellite radar images," *Machine learning,* vol. 30, no. 2, pp. 195-215, 1998.

[10] Y. Li, G. Sun, and Y. Zhu, "Data imbalance problem in text classification," in *2010 Third International Symposium on Information Processing*, 2010: IEEE, pp. 301-305.

[11] S. Wang and X. Yao, "Using class imbalance learning for software defect prediction," *IEEE Transactions on Reliability,* vol. 62, no. 2, pp. 434-443, 2013.

[12] P. Jeatrakul, K. W. Wong, and C. C. Fung, "Classification of imbalanced data by combining the complementary neural network and SMOTE algorithm," in *International Conference on Neural Information Processing*, 2010: Springer, pp. 152-159.

[13] M. Yavaş, A. Güran, and M. Uysal, "Covid-19 Veri Kümesinin SMOTE Tabanlı Örnekleme Yöntemi Uygulanarak Sınıflandırılması," *Avrupa Bilim ve Teknoloji Dergisi,* pp. 258-264, 2020.

[14] N. Çürükoğlu, "Imbalanced Dataset Problem in Classification Algorithms," in *2019 1st International Informatics and Software Engineering Conference (UBMYK)*, 2019: IEEE, pp. 1-5.

[15] P. Hart, "The condensed nearest neighbor rule (corresp.)," *IEEE transactions on information theory,* vol. 14, no. 3, pp. 515-516, 1968.

[16] Ö. Çelik and G. Kaplan, "Yeniden Örnekleme Teknikleri Kullanarak SMS Verisi Üzerinde Metin Sınıflandırma Çalışması," *Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen Bilimleri Dergisi,* vol. 36, no. 3, pp. 434-443, 2020.

[17] I. Tomek, "Two modifications of CNN," 1976.

[18] A. O. Durahim, "Comparison of sampling techniques for imbalanced learning," *Yönetim Bilişim Sistemleri Dergisi,* vol. 2, no. 2, pp. 181-191, 2016.

[19] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," in *Icml*, 1997, vol. 97: Citeseer, pp. 179-186.

[20] D. L. Wilson, "Asymptotic properties of nearest neighbor rules using edited data," *IEEE Transactions on Systems, Man, and Cybernetics,* no. 3, pp. 408-421, 1972.

[21] J. Laurikkala, "Improving identification of difficult small classes by balancing class distribution," in *Conference on Artificial Intelligence in Medicine in Europe*, 2001: Springer, pp. 63-66.

[22] I. Mani and I. Zhang, "kNN approach to unbalanced data distributions: a case study involving information extraction," in *Proceedings of workshop on learning from imbalanced datasets*, 2003, vol. 126: ICML United States.

[23] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of artificial intelligence research,* vol. 16, pp. 321-357, 2002.

[24] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," in *International conference on intelligent computing*, 2005: Springer, pp. 878-887.

[25] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-level-smote: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem," in *Pacific-Asia conference on knowledge discovery and data mining*, 2009: Springer, pp. 475-482.

[26] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, 2008: IEEE, pp. 1322-1328.

[27] M. A. Aydın, "Müşteri kaybı tahmininde sınıf dengesizliği problemi," *Politeknik Dergisi,* pp. 1-1, 2020.

[28] İ. B. Aydilek, "Yazılım hata tahmininde kullanılan metriklerin karar ağaçlarındaki bilgi kazançlarının incelenmesi ve iyileştirilmesi," *Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi,* vol. 24, no. 5, pp. 906-914, 2018.

[29] H. Demir, P. Erdoğmuş, and M. Kekeçoğlu, "Destek Vektör Makineleri, YSA, K-Means ve KNN Kullanarak Arı Türlerinin Sınıflandırılması," *Düzce Üniversitesi Bilim ve Teknoloji Dergisi,* vol. 6, no. 1, pp. 47-67, 2018.

[30] B. Daş and İ. Türkoğlu, "DNA dizilimlerinin sınıflandırılmasında karar ağacı algoritmalarının karşılaştırılması," 2014.

[31] M. F. Amasyalı, B. Diri, and F. Türkoğlu, "Farklı özellik vektörleri ile Türkçe dokümanların yazarlarının belirlenmesi," in *15th Turkish Symposium on Artificial Intelligence and Neural Networks*, 2006.

[32] A. Güran, M. Uysal, and Ö. Doğrusöz, "Destek vektör makineleri parametre optimizasyonunun duygu analizi üzerindeki etkisi," *Dokuz Eylül Üniversitesi Mühendislik Fakültesi Fen ve Mühendislik Dergisi,* vol. 16, no. 48, pp. 86-93, 2014.

[33] L. Tomak and B. Yüksel, "İşlem karakteristik eğrisi analizi ve eğri altında kalan alanların karşılaştırılması," *Journal of Experimental and Clinical Medicine,* vol. 27, no. 2, 2009.