Dengesiz Veri Kümeleri İçin Epileptik Nöbet Tahmini

Epileptic Seizure Prediction for Imbalanced Datasets

Ercan COŞGUN
Elektronik ve Otomasyon Bölümü
Kırklareli Üniversitesi TBMYO, Kırklareli, Türkiye
ercancosgun@klu.edu.tr

Anıl ÇELEBİ ve M. Kemal GÜLLÜ
Elektronik Haberleşme Mühendisliği
Kocaeli Üniversitesi Müh. Fakültesi, Kocaeli, Türkiye
anilcelebi@kocaeli.edu.tr, kemalg@kocaeli.edu.tr

***Özetçe*— Bu çalışmada dengesiz veri setlerinin sınıflandırılmasında kullanılan yöntemler epilepsi hastalarından alınan EEG işaretlerine uygulanıp, epileptik nöbet tahmini yapılmıştır. Öncelikle örnek azaltma, örnek çoğaltma, sentetik örnek elde etme yöntemleri kullanılarak veri seti dengeli hale getirilip Destek Vektör Makineleri ile sınıflandırılmıştır. Sonrasında, veri seti dengeli hale getirilmeden, doğrudan RUSBoost sınıflandırıcı kullanılarak sınıflandırılmıştır. Sınıflandırma sonuçları, farklı ölçütler ile karşılaştırılmış ve yöntemlerin üstünlükleri ve zayıflıkları değerlendirilmiştir.**

Anahtar Kelimeler — epileptik nöbet tahmini; dengesiz veri seti; rusboost sınıflandırıcı.

***Abstract*— In this study, the methods used in the classification of imbalanced data sets were applied to EEG signals obtained from epilepsy patients and epileptic seizures were estimated. Firstly, the data set was balanced by using under-sampling, over-sampling, and synthetic minority over-sampling technique and classified with Support Vector Machines. Then, the data set was classified using the Rusboost classifier without balancing. Classification results were compared with different criteria and the advantages and disadvantage of the methods were evaluated.**

Keywords — epileptic seizure prediction; imbalanced dataset; rusboost Classifier.

I. GİRİŞ

Makine öğrenmesi sistemin geçmişteki tecrübelerini kullanarak bir model oluşturan ve gelecekteki karşılaşacağı durumlar karşısında bir tahminde bulunan yapay zekâ alanıdır. Temel olarak kendi kendine karar verme mekanizmasına dayanan, makine öğrenmesinde iki temel yöntem kullanılmaktadır; denetimli öğrenme ve denetimsiz öğrenme. Denetimli öğrenmede sınıflara ait etiket değerleri verilirken, denetimsiz öğrenmede etiket bilgileri verilmemektedir. Etiketler olmadığı için veriyi değişkenler arasındaki ilişkilere dayalı olarak kümeleyerek bir model oluşturur. Tespit ve tahmin uygulamalarında genellikle denetimli öğrenme yöntemi kullanılmaktadır.

Denetimli öğrenmede kullanılan algoritma ve algoritma parametreleri kadar kullanılan veri setinin yapısın da oldukça önemlidir. Özellikle tahmin sistemi çalışmalarında veri setine ait gözlemler eşit dağılmayabilir. Bu durum dengesiz veri seti (imbalanced dataset) olarak tanımlanmaktadır. Dengesiz veri seti ile bir önlem alınmadan sınıflandırma yapıldığında bir sınıfa ait örnek sayısının daha fazla olmasından dolayı modelde bir ön yargı oluşacaktır. Ön yargı oluşmasıyla daha az sınıf türleri için başarısız bir sınıflandırma modeli oluşmaktadır.

Epilepsi beyin hücrelerinin elektriksel aktivitesinin anormal olarak aniden bozulmasıyla tuhaf hareketler, duyusal bozukluklar ve bilinç kapanması ile ortaya çıkan, ani olarak görülen sinir sistemi hastalığıdır. Epilepsi tahmin ve tespit çalışmalarında EEG (Elektroensefalografi) çizelgesi kullanılmaktadır. Epilepsi hastalarında EEG işareti Preiktal (Nöbet öncesi), İktal (Nöbet anı), postiktal (Nöbet Sonrası) ve interiktal (Nöbetler arası) olmak üzere dört sınıfa ayrılmaktadır. Tahmin sistemlerinde daha çok preiktal ve interiktal evre öngörülmeye çalışılmaktadır. Preiktal evre nöbet anından yaklaşık bir saat öncesine kadar alınabilen evre olarak tanımlanmaktadır. Bu durumda preiktal evreye ait sınıf sayısı interiktal evreye göre oldukça düşük kalmaktadır. Bu tür durumlarda uygun sınıflandırıcı seçimi ya da dengesiz veri seti ile kullanılan yöntemler tercih edilmektedir.

Literatürde epilepsi nöbet tahmini ve tespiti üzerine dengesiz veri setleri kullanılarak bazı çalışmalar yapılmıştır. Teixeira ve arkadaşlarının yapmış olduğu çalışmada interiktal evreye ait örnek sayısı rastgele seçilerek diğer evrelere ait örnek sayısının toplamına eşitlenmiştir. Aynı çalışmada destek vektör makinesi (DVM) kullanılarak maliyet parametresi değiştirilmiş fakat bu tekniğin sınıf dengesizliğinin çok kuvvetli olduğu durumlarda çalışmadığından bahsedilmiştir [1]. Truong ve arkadaşlarının önerdiği çalışmada eğitim aşamasında üst üste gelen örnekleme tekniği kullanılarak preiktal evreye ait yeni örnekler üretilmektedir. Böylece örnek çoğaltma yöntemi kullanılarak interiktal ve preiktal sınıf sayıları eşitlenmiştir [2]. Alickovic ve arkadaşlarının yapmış olduğu çalışmada ise DVM ile örnek azaltma yöntemi kullanılarak sınıflandırma yapılmıştır [3]. Karumuri ve arkadaşları Sentetik Azınlık Örnekleme Tekniği (SMOTE) kullanarak veri setini dengeli hale getirmiştir. Bu teknik azınlık sınıfının birbirine en yakın ya da belirlenen sayıdaki komşusu arasında her öznitelik için sentetik yeni örnekler üretmektedir. Bu sayede azınlık sınıfları, çoğunluk sınıfları sayısına eşitlenmiştir [4]. Diğer çalışmalardan farklı olarak Solaija ve arkadaşlarının yapmış olduğu çalışmada ise veri seti dengeli hale getirilmeden RusBoost sınıflandırıcı kullanılarak

978-1-7281-2420-9/19/$31.00 ©2019 IEEE

sınıflandırılmıştır. Kullanılan bu sınıflandırıcı dengesiz veri setlerinde daha faydalı sonuç verdiğinden bahsedilmektedir [5]. Aynı şekilde Amin ve Kamboh da RusBoost sınıflandırıcıya ek SMOTEBoost olarak tanımladıkları sınıflandırıcı yöntemi ile melez bir sistem önermiştir [6].

Bu çalışmada ise epileptik nöbet kestirimi için örnek çoğaltma, örnek azaltma, sentetik azınlık örnekleme tekniği yöntemleriyle veriler dengeli hale getirilip DVM ve dengeli hale getirilmeden RusBoost algoritması kullanılarak sınıflandırılmıştır. Bu sınıflandırıcıların performans analizleri yapılmıştır. Performans analiz sonuçları değerlendirilerek dengesiz veri seti ile epilepsi nöbet kestirimi için hangi yöntemin daha uygun olacağı önerilmiştir.

II. MATERYAL VE METOD

A. Veri seti

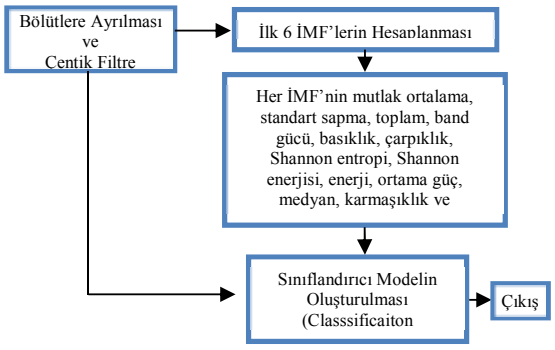
Çalışmada kullanılan EEG verileri "The European Epilepsy Database" tarafından 256 Hz örnekleme hızı ile kaydedilmiştir [7]. Toplam 30 epilepsi hastasına ait 29 kanal EEG verisi içermektedir. Yapılan çalışmada Hasta_1'e ve Hasta_4'e ait verilerden saçsız deriye denk gelen FP1 kanalından alınan EEG sinyali kullanılmıştır. Hastalar rastgele seçilmiştir. Tablo1'de EEG verileri hakkında bilgiler verilmiştir.

TABLO I. HASTA BILGILERI

	Toplam Nöbet Sayısı	Eğitim için Ayrılan Nöbet sayısı	Test için Ayrılan Nöbet sayısı
Hasta_1	11	5	6
Hasta_4	5	3	2

B. Sistem mimarisi

Tasarlanan sistemin mimarisi Şekil 1'de görülmektedir.



Şekil. 1.Yapılan çalışmanın blok diyagramı

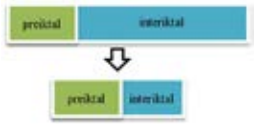
Tasarlanan sistemde eğitim ve test için ayrılan veriler 10 saniyelik bölütlere ayrılarak şebekeden kaynaklı gürültülerin yok edilmesi için 50 Hz'lik Çentik (Notch) filteresinden geçirilmiştir. Daha sonra ampirik kip ayrışımı (AKA) kullanılarak sinyalin ilk 6 içsel mod fonksiyonu (İMF) elde edilmiştir. AKA doğrusal ve durağan olmayan zaman serileri analizi için kullanılan bir yöntemdir [8]. Bu yöntem kullanılarak sinyaller içsel mod fonksiyonu(İMF) adı verilen alt bileşenlere ayrıştırılmaktadır. Her bir İMF'nin Şekil 1'de görüldüğü gibi literatürde sıklıkla kullanılan öznitelikler çıkartılmıştır [1,3,9,10]. Son olarak dört yöntem için ayrı ayrı veri seti oluşturularak Mathworks firmasına ait "Matlab Classification Learner" aracı ile sınıflandırıcı modeller oluşturularak test edilmiştir. Bu araç kullanılarak sınıflandırıcı modellin oluşturulması, öznitelik seçimi ve sonuçların değerlendirilmesi yapılabilmektedir.

III. SINIFLANDIRICI MODELLERİN OLUŞTURULMASI

Yapılan çalışmada dört farklı sınıflandırma modeli oluşturulmuştur.

A. Örnek sayısının azaltılması (SVM_r)

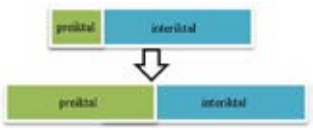
Örnek sayısının azaltılması sınıf dağılımını değiştirmeyi amaçlayan popüler bir yeniden örnekleme stratejisidir. Sınıflandırıcıları eğitmeden sistemi dengeli hale getirmektedir. Bu işlem rastgele yapılacağı gibi, belirlenen bir stratejide de yapılabilmektedir. Yapılan çalışmada Hasta_1 ve Hasta_4'e ait interiktal ve preiktal sınıflar Şekil 2'de olduğu gibi dengeli hale getirilmiştir. İnteriktal sınıfına ait örnekler rastgele seçilmiştir.



Şekil. 2.Örnek sayısının azaltılması

B. Örnek sayısının çoğaltılması (SVM_e)

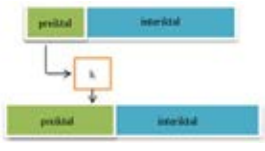
Örnek sayısının çoğaltılması sınıflar arası dengeyi sağlamak amacıyla sayıca az olan sınıfa ait örneklerin tekrar kopyalanarak, sayıca çok olan sınıfa ait örneklerin sayısına eşitleme işlemidir. Şekil 3'de görüldüğü gibi preiktal sınıfına ait örnekler interiktal örnek sayısına eşitlenmiştir.



Şekil. 3.Örnek sayısının çoğaltılması

C. Sentetik azınlık örnekleme tekniği (SVM_s)

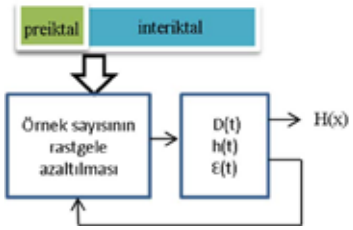
Bu yöntemde her azınlık sınıfına ait örneğin k kadar komşusuna bakılarak yapay örnekler üretilir. Bu sayede Şekil 4'deki gibi preiktal sınıfına ait örnek sayısı interiktal sınıfına eşitlenir. Bu işlemin örnek sayısını çoğaltılmasından farkı azınlık sınıflarındaki örneklerin hepsinin kopyalanması yerine sadece yakın komşularına bakılarak yeni yapay örnekler oluşturulmasıdır.



Şekil. 4.Sentetik azınlık örnekleme tekniği

D. RUSBoost Sınıflandırıcı (RusBoost)

RUSBoost, sınıflar arası dengesizlik oranı çok fazla olduğu zaman tercih edilen sınıflandırıcıdır [11]. Bu yöntem alt örnekleme kullanarak istenilen denge oluşana kadar sayıca çok olan sınıftan örnekleri rastgele kaldırmaktadır.



Şekil. 5.RusBoost sınıflandırıcı

RusBoost'un genel işleyişi Şekil 5'de gösterilmektedir. Burada t bir ile T (topluluk sınıflandırıcısının sayısı) arası iterasyon sayısını; h(t), t' inci iterasyonda eğitilmiş zayıf hipotezleri ve D(t) örneğin ağırlığını temsil etmektedir. Zayıf hipotezler T defa döngüye girerek eğitilir. Örnek sayısı çoğunluk sınıf örneklerinden, azınlık örnek sayısına eşit oluncaya kadar rastgele azaltılır. Ağırlık dağılımına göre Ɛ(t) (hata) hesaplanır. Sonuç olarak H(x) hipotezi elde edilir. H(x) zayıf hipotezlerin ağırlıklı oyunu döndürmektedir [12].

IV. SINIFLANDIRICI PERFORMANS PARAMETRELERİ

Veri dengesizliği çok büyük olduğunda, sınıflandırma başarımı için sadece doğruluğa bakmak yetersiz olacaktır. Çünkü doğruluk oldukça yüksek çıkacaktır. Bu durumda doğruluktan başka diğer performans parametrelerine bakılması gerekmektedir. Sınıflandırma algoritmasında kullanılan performans parametreleri epilepsi nöbet tahmini için uyarlanarak aşağıda verilmiştir.

A. Karmaşıklık Matrisi

Karmaşıklık matrisi sınıflandırma sonucunun doğruluğu hakkında bilgi veren ölçüm aracıdır.

TABLO II. KARMAŞIKLIK MATRISI

		Tahmin	
Gerçek		Doğru Pozitif (DP)	Yanlış Negatif (YN)
		Yanlış Pozitif (YP)	Doğru Negatif (DN)

Tablo 2'de karmaşıklık matrisi verilmiştir. Bu tabloya göre gerçek etiketler ile sınıflandırma sonuçları karşılaştırılmaktadır. Tablo 2' deki hücrelerde bulunan etiketlerin anlamları aşağıda verilmiştir.

DP: Gerçekte preiktal, sınıflandırıcı sonucu preiktal.

DN: Gerçekte interiktal, sınıflandırıcı sonucu interiktal.

YN: Gerçekte preiktali sınıflandırıcı sonucu interiktal.

YP: Gerçekte interiktal, sınıflandırıcı sonucu preiktal

Makine öğrenmesindeki diğer tüm performans değerleri bu tabloya göre hesaplanmaktadır.

B. Doğruluk (D)

Denklem (1)' de verilen doğruluk, preiktal ve interiktal evrelerinin toplam hangi doğrulukta sınıflandırdığını gösterir.

$$D = (DP + DN) / (DP + YN + YP + DN) \quad (1)$$

C. Doğru pozitif oranı (DPO)

Sadece preiktal evreleri hangi oranda doğru olarak sınıflandığı bilgisini vermektedir. Bu değer epilepsi nöbet tahmin sınıflandırmasında düşük çıkabilmektedir. Çünkü eğitim veri setinde hastanın tüm preiktal evrelerine ait örnekler bulunmamaktadır. Bu değer (2)'deki gibi hesaplanmaktadır.

$$DPO = (DP / (DP + YN)) \quad (2)$$

D. Doğru negatif oranı (DNO)

Sadece interiktal evreleri hangi oranda doğru olarak sınıflandığı bilgisini vermektedir. Bu değer (3)'deki gibi hesaplanmaktadır.

$$DNO = (DN / (DN + YP)) \quad (3)$$

E. Yanlış pozitif oranı (YPO)

Denklem (4)'de verilen YPO, hangi oranda interiktal evreleri preiktal olarak sınıflandırdığını gösterir. Epilepsi nöbet tahmin sisteminde bu oranın düşük çıkması beklenmektedir. Çünkü böyle tasarlanan bir sistemin çok fazla yanlış alarm vermesi istenememektedir. Yapılan bazı çalışmalarda son işlemci olarak "Fring power" algoritması kullanılarak yanlış alarmların sayısı azaltılmaktadır [1,9].

$$YPO = (YP / (YP + DN)) \quad (4)$$

F. Yanlış Negatif oranı (YNO)

Hangi oranda preiktal evreleri interiktal olarak sınıflandırdığını gösterir. YNO'nun nasıl hesaplandığı (5)' de gösterilmiştir.

$$YNO = (YN / (YN + DP)) \quad (5)$$

G. Pozitif kesinlik oranı (PKO)

Denklem (6)' da preiktal evre olarak sınıflandırılan evrelerin, gerçekten kaçının preiktal evre olduğunu gösterir.

$$PKO = (DP / (DP + YP)) \quad (6)$$

H. Ağırlıklı ortalama duyarlığı (AOD)

Denklem (7)'de gösterilen AOD sınıflandırıcı modelin iyi olup olmadığı hakkında bilgi vermektedir.

$$AOD = (DPO + DNO) / 2 \quad (7)$$

I. F Ölçütü (F)

Denklem (8)'deki bağıntı ise duyarlılık ile kesinlik arasındaki bağıntısı vermektedir.

$$F = (2 \times DPO \times PKO) / (DPO + PKO) \quad (8)$$

V. SONUÇLAR

Bu çalışmada 2 hastaya ait EEG kayıtlarından elde edilen veriler ile epilepsi tahmin sistemi üzerinde dengesiz veri setlerinde kullanılan yöntemler uygulanarak sınıflandırıcı modeller oluşturulmuştur. Bu sınıflandırıcı modellere ait performans parametreleri hesaplanmıştır. Hasta_1'e ait başarım değerleri Tablo 3'de, Hasta_4'e ait değerler ise Tablo 4'de verilmiştir.

TABLO III. HASTA_1'E AIT PERFORMANS DEĞERLERI

Yöntem	D	DPO	DNO	YPO	YNO	PKO	AOD	F
SVM_r	0,50	0,44	0,50	0,49	0,55	0,03	0,47	0,06
SVM_e	0,64	0,22	0,66	0,33	0,77	0,02	0,44	0,04
SVM_s	0,71	0,18	0,73	0,26	0,81	0,02	0,45	0,04
RusBoost	0,75	0,59	0,76	**0,23**	**0,40**	0,08	**0,67**	0,15

TABLO IV. HASTA_4'E AIT PERFORMANS DEĞERLERI

Yöntem	D	DPO	DNO	YPO	YNO	PKO	AOD	F
SVM_r	0,91	0,95	0,90	0,09	0,04	0,32	0,92	0,47
SVM_e	0,96	0,66	0,98	**0,01**	0,33	0,63	0,82	0,64
SVM_s	0.90	0,57	0,97	0,02	0,42	0,54	0,77	0,55
RusBoost	0,89	0,97	0,89	0,10	**0,02**	0,30	**0,93**	1,20

Sınıflandırıcı parametrelerinden F ölçütü sınıf dengesizliği probleminden etkilendiği için RusBoost modelinin F değeri ile diğer modellerin F değeri karşılaştıramamaktadır. Çünkü RusBoost modeli dışındaki modeller dengeli veri seti ile eğitilmiştir. Bu durumdaki sınıflandırıcı parametrelerini karşılaştırmak için öncelikle AOD' e bakılması gerekmektedir. Fakat AOD, YPO hakkında bilgi vermemektedir. Bu yüzden AOD' e ek olarak YPO değerine de bakılmaktadır. YPO değeri ise olabildiğince düşük çıkması gerekmektedir.

Tablo 3'e bakıldığında RusBoost modeline ait AOD değeri yüksek çıktığı görülmektedir. Aynı zamanda YPO değerleri de diğerlerine göre düşük çıktığı gözlemlenmektedir.

Tablo 4'de ise RusBoost modeline ait AOD değerleri yüksek çıkmıştır. Fakat YPO değeri SVM_e modelinde daha düşük çıkmıştır. Aynı zamanda hem Hasta_1'in ve Hasta_4' ün YNO değeri de RusBoost modelinde daha düşük çıkmıştır.

VI. TARTIŞMA

Yapılan çalışmada SVM_e ve SVM_s sınıflandırıcı modeli kullanıldığında örnek sayısı artığı için eğitim süresini artırmaktadır. SVM_r modelinde ise örnek azaltma işlemi yapıldığı için interiktal sınıfına ait yararlı örnekler yok edilmektedir. RusBoost modeline bakıldığında ise eğitimde kullanılacak veri setini dengeli hale getirmeden sınıflandırabilmektedir. Ayrıca alt örneklerden kaynaklanan bilgi kaybını da önlemektedir. Sınıflandırma performans parametrelerine bakıldığında her iki hasta içinde RusBoost sınıflandırıcı modeli kullanılması daha uygun olduğu önerilmektedir.

BİLGİLENDİRME

Bu çalışma Kocaeli Üniversitesi BAPKO (2018/063) tarafından desteklenmiştir.

KAYNAKLAR

[1] Teixeira, C. A., Direito, B., Bandarabadi, M., Le Van Quyen, M., Valderrama, M., Schelter, B., ... & Dourado, A. (2014). *Epileptic Seizure Predictors Based on Computational Intelligence Techniques: A comparative study with 278 patients. Computer methods and programs in biomedicine*, 114(3), 324-336.

[2] Truong, N. D., Nguyen, A. D., Kuhlmann, L., Bonyadi, M. R., Yang, J., Ippolito, S., & Kavehei, O. (2018). *Convolutional neural networks for seizure prediction using intracranial and scalp electroencephalogram. Neural Networks*, 105, 104-111.

[3] Alickovic, E., Kevric, J., & Subasi, A. (2018*). Performance evaluation of empirical mode decomposition, discrete wavelet transform, and wavelet packed decomposition for automated epileptic seizure detection and prediction. Biomedical Signal Processing and Control*, 39, 94-102.

[4] Karumuri, B. K., Vlachos, I., Liu, R., Adkinson, J. A., & Iasemidis, L. (2016, March). *Classification of pre-ictal and interictal periods based on EEG frequency features in epilepsy. In 2016 32nd Southern Biomedical Engineering Conference (SBEC)* (pp. 9-10). IEEE

[5] Solaija, M. S. J., Saleem, S., Khurshid, K., Hassan, S. A., & Kamboh, A. M. (2018). *Dynamic mode decomposition based epileptic seizure detection from scalp EEG. IEEE Access*, 6, 38683-38692.

[6] Amin, S., & Kamboh, A. M. (2016, September). *A robust approach towards epileptic seizure detection. In 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP)* (pp. 1-6). IEEE.

[7] www.epilepsiae.eu/project_outputs/european_database_on_epilepsy

[8] Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., & Liu, H. H. (1998). *The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis. Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 454(1971), 903-995.

[9] Direito, B., Teixeira, C. A., Sales, F., Castelo-Branco, M., & Dourado, A. (2017). *A realistic seizure prediction study based on multiclass SVM. International Journal of Neural Systems*, 27(03), 1750006.

[10] Rasekhi, J., Mollaei, M. R. K., Bandarabadi, M., Teixeira, C. A., & Dourado, A. (2013). *Preprocessing effects of 22 linear univariate features on the performance of seizure prediction methods. Journal of Neuroscience Methods*, 217(1-2), 9-16.

[11] Tomak, Ö., Güler, B., Tüfekçi, A., & Yanmaz, K. Dalgacık *Dönüşümü ve RUS Geliştirilmiş Ağaç Kullanarak Otomatik Aritmi Tespiti. Karadeniz Fen Bilimleri Dergisi*, 8(2), 1-9.

[12] Seiffert, C., Khoshgoftaar, T. M., Van Hulse, J., & Napolitano, A. (2010). *RUSBoost: A hybrid approach to alleviating class imbalance. IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(1), 185-197.