Pamukkale Univ Muh Bilim Derg, 26(2), 385-391, 2020

Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi

Pamukkale University Journal of Engineering Sciences

Yenidoğan kuzularda bilgisayar destekli tanı

Computer-aided diagnosis in neonatal lambs

Pınar CİHAN[1], Oya KALIPSIZ[2], Erhan GÖKÇE[3]*

[1]Bilgisayar Mühendisliği Bölümü, Mühendislik Fakültesi, Namık Kemal Üniversitesi, Tekirdağ, Türkiye.
pkaya@nku.edu.tr

[2]Bilgisayar Mühendisliği Bölümü, Elektrik-Elektronik Fakültesi, Yıldız Teknik Üniversitesi, İstanbul, Türkiye.
kalipsiz@yildiz.edu.tr

[3]İç Hastalıkları Anabilim Dalı, Veteriner Fakültesi, Kafkas Üniversitesi, Kars, Türkiye.
erhangokce36@hotmail.com

Geliş Tarihi/Received: 29.11.2018 Düzeltme Tarihi/Revision: 26.03.2019 doi: 10.5505/pajes.2019.51447
Kabul Tarihi/Accepted: 24.04.2019 Araştırma Makalesi/Research Article

Öz

Ülkemizde küçükbaş hayvan sayısı her geçen gün çeşitli sebeplerden dolayı azalmaktadır. Küçükbaş hayvan sayısının azalmasına paralel olarak, hayvansal üretimde de önemli azalmalar görülmektedir. Küçükbaş hayvan sayısının azalmasını önlemenin bir yolu da hastalıklarla ilgili tahmin ve analizlerin başarılı bir şekilde yapılabilmesidir. Makine öğrenmesi ile yapılan bilgisayar destekli tanı çalışmaları sayesinde, sağlık hizmetlerinin kalitesi artarken sağlık sektöründeki maliyetler azalmaktadır. Bu çalışmanın amacı makine öğrenmesi yöntemleri ile kuzularda erken hastalık teşhisi yapmaktır. Bunun için çalışmada karar ağaçları, saf bayes, k-en yakın komşu, yapay sinir ağları ve rassal orman yöntemleri kullanılmıştır. Bu sınıflandırma yöntemlerinin performansları doğruluk, dengeli doğruluk, seçicilik, duyarlılık, F-ölçütü, kappa ve ROC eğrisi altında kalan alan (AUC) ölçütleri ile analiz edilmiştir. Çalışma sonucunda bilgisayar destekli tanı için Saf bayes yöntemi diğer yöntemlerden daha başarılı sonuçlar üretmiştir. Basit ve uygulaması kolay olan Saf bayes yöntemin diğer karmaşık yöntemlerden daha başarılı sonuçlar elde etmesi oldukça önemlidir.

Anahtar kelimeler: Bilgisayar destekli tanı, Sınıflandırma, Saf bayes, Küçükbaş hayvan.

Abstract

In our country, the number of small ruminant animals is decreasing day by day due to various reasons. In parallel with the decrease in the number of small ruminants, significant decreases are seen in animal production. One way to prevent the reduction in the number of small ruminants is to be able to make successful predictions and analysis related to the diagnosis. Thanks to computer-aided diagnostic studies performed with machine learning, the quality of health services increases while the costs of the health sector decrease. The aim of this study is to perform computer aided diagnosis in neonatal lambs using machine learning methods. Hence in study, decision tree, naive bayes, k-nearest neighbors, artificial neural networks and random forest methods were used. The performances of these classification methods were analyzed with accuracy, balanced accuracy, specifity, recall, F-measure, kappa and area under the ROC curve (AUC) criteria. As a result of the study, the Naive bayes method more successful results than other methods for computer aided diagnosis produced. It is very important that, the Naive bayes method is simple and easy to apply, achieves more successful results than other complex methods.

Keywords: Computer-aided diagnosis, Classification, Naive bayes, Small ruminant animal.

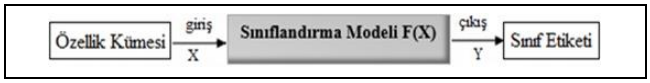
1 Giriş

Ülkemizde küçükbaş hayvan sayısı ve hayvansal ürünlerdeki azalma özellikle kırsal kesimin daha da yoksullaşmasına neden olmaktadır. Yoksullaşmanın önüne geçebilmek için öncelikle bu sektördeki karlılığı veya verim performansını arttırmak şarttır. Bu da koyunculuğa talebi arttırmayı gerektirmektedir. Koyunculuğa talebi arttırmak için de hastalıklar ve ölüm oranlarının azaltılması gerekmektedir. Bu nedenle hayvanın hastalık durumu ile ilgili tahmin ve analizlerin yapılması önemli bir hususdur.

Günümüzde veri madenciliği (VM) yöntemlerinden, birçok alanda özellikle tıp alanında sıklıkla yararlanılmaktadır[1]-[3]. VM, en objektif ve optimum çözümleri kullanarak hekimlerin en doğru ve güncel bilgiye ulaşmasını sağlayacak bir karar destek aracıdır. Bu karar destek aracı sayesinde hekim kararı desteklenebilir, ön görülemeyen bilgiler açığa çıkarılabilir veya yeni bir örneğin sınıf bilgisi tahmin edilebilir[4],[5].

Veri madenciliğinde, verinin içerdiği ortak özelliklere göre ayrıştırılması işlemi sınıflandırma olarak adlandırılmaktadır. Sınıflandırma, sınıfı belli olan örneklerden yola çıkarak, sınıfı belli olmayan örneklerin sınıfını tahmin etmek için kullanılan VM modelidir [6].

Veri madenciliği modelleri, tahmin edici ve tanımlayıcı olmak üzere iki kategoriye ayrılır. Tahmin edici algoritmalar, hedef değişken (kesikli veya sürekli) türüne göre sınıflandırma ve regresyon olarak ayrılır. Sınıflandırma ise, eğiticili ve eğiticisiz olmak üzere ikiye ayrılmaktadır. Eğiticili sınıflandırma; sınıf bilgisi bilinen verileri kullanılarak elde edilen modelle, sınıf bilgisi bilinmeyen yeni verilerin sınıflandırılmasıdır [7]. Sınıflandırma süreci Şekil 1'de gösterilmiştir.



Şekil 1. Sınıflandırma süreci.

Figure 1. Classification process.

*Yazışılan yazar/Corresponding author

Sınıflandırma işleminin ilk basamağında eğitim verilerinin kullanılmasıyla tahmin için kullanılacak bir model oluşturulur. Sonraki adımında ise elde edilen model kullanılarak daha önce hiç görülmemiş (sınıfı belli olmayan veriler) üzerinde uygulanarak sınıflar tahmin edilir [6]. Bu işlemler eğitim ve test adımları olarak da adlandırılmaktadır. Eğiticisiz sınıflandırma yani diğer bir adıyla kümeleme ise sınıf bilgisi bilinmeyen verilerin karşılaştırma yoluyla gruplandırılması işlemine dayanır.

VM yöntemleri tıp alanında yaygınlıkla teşhis, tedavi veya hekim kararını destekleyici bir sistem olarak yaygınlıkla kullanılmasına karşın en az insan sağlığı kadar önemli olan hayvan sağlığında bu etkili yöntemden yararlanma son yıllarda ivme kazanmıştır [8]. Veterinerlik alanında yapılmış çalışmalarda VM yöntemlerinden başarılı sonuçlar elde edilmiş olup, veterinerlik alanında etkili bir şekilde uygulanabileceği bildirilmiştir [8].

Veterinerlik alanında; kızgınlık tespiti[9],[10], sığır sıcaklık stres tahmini [11], süt verimi tahmini [12], süt gübresindeki besin içeriğinin tahmin etme [13],[14], çiğ süt kalitesinin değerlendirilmesi[15], döllüğe etki eden faktörlerin belirlenmesi [16], koyunları vücut ölçülerine göre sınıflandırma [17], çiftliklerde risk sınıflandırılması [18], hastalık teşhisi ve risk faktörlerinin belirlenmesi [19] gibi çeşitli konularda VM yöntemlerinden yararlanılmıştır.

Bu çalışmada neonatal (yenidoğan) kuzularda bilgisayar destekli tanı yapılmıştır. Literatürde böyle bir çalışma yapılmamıştır [8]. Hasta kuzuların otomatik sınıflandırılması sayesinde veteriner hekime kararında yardımcı olacak bir sitem geliştirilmiş olacaktır. Ayrıca hasta kuzuların erken tespiti sayesinde önlemler alınarak, ölümlerin önüne geçilebilecektir. Hasta ve sağlıklı kuzu ayrımı gerçekleştirilirken eğiticili sınıflandırma yöntemlerinden karar ağaçları, saf bayes, k-en yakın komşu, sinir ağları ve rassal orman algoritmaları kullanılmıştır. Kullanılan sınıflandırma yöntemlerinin performansları duyarlılık, seçicilik, F-ölçümü, doğruluk, dengeli doğruluk ve ROC eğrisi altında kalan alan (AUC) ölçütlerine göre karşılaştırılarak neonatal kuzularda hastalık sınıflandırılmasında en başarılı yöntem belirlenmiştir.

2 Materyal ve yöntem

Bu çalışmada neonatal kuzularda bilgisayar destekli tanı işleminin gerçekleştirilmesi için Karar Ağaçları (KA, Decision Tree-DT), Saf Bayes (SB, Naïve Bayes - NB), K- en yakın komşu (K-EYK, K-Nearest Neighborhood-KNN), Sinir Ağları (SA, Neural Network-NN) ve Rassal Orman (RO, Random Forest-RF) sınıflandırma yöntemleri kullanılmıştır. Bu yöntemler farklı metrikler ile karşılaştırılarak, sağlıklı ve hasta kuzuların sistem tarafında otomatik olarak başarılı bir şekilde sınıflandırmasını sağlayacak en başarılı yöntem belirlenmeye çalışılmıştır.

Literatürde genellikle veri setinin 2/3'ü eğitim, 1/3'ü test olarak kullanılmaktadır [20]. Çalışmada veri setinin %70'i eğitim kümesi, %15'i test kümesi ve %15'i doğrulama kümesi olarak kullanılmıştır. Ayrıca her bir farklı eğitim kümesi için 10-kat çapraz geçerlilik (CV, cross-validation) [21] uygulanmıştır. Tüm işlemler R [22] programlama dili kullanılarak gerçekleştirilmiştir.

2.1 Sınıflandırma yöntemleri

Karar Ağaçları (KA, Decision Tree – DT) algoritması, geçmiş veriye dayanarak yeni verilerin hangi sınıfa ait olduğuna, kurallar çıkartarak karar vermektedir. Ağaç, dallar ve yapraklardan oluşur. Eğer yaprak artık dallara ayrılmıyorsa o yaprağa "karar düğümü" denir. Tüm yapraklar karar düğümü olana kadar ya da o yaprağa ait veri kalmayana kadar dallara ayrılmaya devam eder [23],[24]. ID3, C4.5, CART gibi farklı algoritmalar kullanılarak karar ağacı oluşturulabilir. C4.5 algoritması, ID3 algoritmasının geliştirilmesi ve eksiklerinin gidermesi sonucunda oluşturulmuştur. ID3 algoritması kategorik değerler için çalışan bir algoritma iken C4.5 algoritması sayısal değerleri de analize dahil ederek, ağaç yapısına katmaktadır. Bu yöntemi uygulamak için R'da 'caret' paketindeki 'J48' fonksiyonu kullanılmıştır.

Saf Bayes (SB, Naïve Bayes-NB) algoritması, istatistikteki Bayes teoremine dayanmaktadır. Bu teorem; belirsizlik taşıyan herhangi bir durumun modelinin oluşturularak, bu durumla ilgili evrensel doğrular ve gerçekçi gözlemler doğrultusunda belli sonuçlar elde edilmesine olanak sağlar. Bayes kuralına göre sınıflandırılacak örneğe, en yüksek olasılıkla benzerlik gösteren sınıf seçimi ile yapılır. Bu seçim hesaplanırken önsel olasılıklardan yararlanılır. Bu yöntemi uygulamak için R'da 'caret' paketindeki 'nb' fonksiyonu kullanılmıştır.

K-En Yakın Komşu (k-EYK, K-Nearest Neighborhood-KNN) algoritması, hangi sınıfa ait olduğu bilinmeyen örneğe sınıf etiketi vermek için, eğitim kümesindeki örneklerin bu örneğe olan uzaklık ölçüsü hesaplanır. Kendisine en yakın örnekler (mesafe ölçüsü en küçük olan örnekler) seçilerek bu örneğin sınıf bilgisi yeni gelen örneğe verilir. Buradaki "k" değeri en yakın kaç komşuya bakılacağını yani komşu sayısını belirtmektedir. Sınıf etiketinin çoğunluk seçimiyle belirlenmesinden dolayı "k" değeri genellikle 3, 5 veya 7 gibi tek sayıda örnek seçilir. K-EYK algoritmasında komşular arasındaki uzaklık genellikle Öklid uzaklığı ile bulunmakla birlikte Mahalanobis [25], Hamming [26], Manhattan [27], Minkowski [28] gibi uzaklık ölçütleri kullanılabilir. Bu yöntemi uygulamak için R'da 'caret' paketindeki 'knn' fonksiyonu kullanılmıştır.

Yapay Sinir Ağı (YSA, Artificial Neural Network-ANN), öğrenebilen bir algoritma olup nöronların çalışma prensibini modellemektedir. Yinelemeli olarak ileriye ya da geriye yönelik besleme alabilen iki tür yapısı vardır. Sinir ağı eğitilirken giriş verilerine karşılık çıkış verileri alınır. Bu değer gerçek değerlerle karşılaştırılır ve ağın içerisindeki nöron fonksiyonlarının bu sonuçtaki hata miktarına göre ayarlanması sağlanır. Bu şekilde birçok değer ağa verilir ve ağın eldeki verinin yapısının öğrenilmesi sağlanır. Öğrenme işlemi tamamlandıktan sonra sinir ağı kullanıma hazır hale gelir. Bu yöntemi uygulamak için R'da 'caret' paketindeki 'nn' fonksiyonu kullanılmıştır.

Rassal Orman (RO, Random Forest-RF), birçok karar ağacının birleştirilmesi ile oluşan bir topluluk yöntemidir [29]. Topluluk öğrenme yöntemlerinde (Ensemble Learning) birden çok sınıflayıcının ortaya koyduğu sonuçlar bir araya getirilerek, topluluk adına tek bir karar verilmektedir. Ormanındaki her karar ağacı, orijinal veri setinden bootstrap (yeniden yerleştirilerek örneklenen) tekniği ile farklı örneklemler seçilerek oluşturulur ve yine rastgele torbalama (bagging) mekanizması ile seçilen bir özellik kümesi ile eğitilir [30]. Daha sonra, birbirinden farklı çok sayıda bireysel ağaç tarafından verilen kararlar oylamaya tabi tutar ve oylama sonucunda en çok oyu alan sınıfı topluluğun (komitenin) sınıf tahmini olarak sunar. Bu yöntemi uygulamak için R'da 'caret' paketindeki 'rf' fonksiyonu kullanılmıştır.

2.2 Sınıflandırma algoritmalarının başarısını test etme ölçütleri

Bilgisayar destekli tanıda, sınıflandırma algoritmalarından elde edilen sonuçların kıyaslanabilmesi için nesnel değerlendirme ölçütlerine ihtiyaç duyulmaktadır. Bu ölçütler modelin ne derece başarılı olduğunu değerlendirmek açısından önem arz etmektedir. Sınıflandırma sonucu elde edilen sonuçlar karışıklık matrisi (confusion matrix) ile ifade edilmektedir. Karışıklık matrisinin sütunları örneklerin gerçek değerlerine satırları ise sınıflandırma sonucu elde edilen sonuçlara karşılık gelmektedir. Bu çalışmada sınıflandırma başarıları değerlendirilirken duyarlılık (recall), seçicilik (specifity), F-ölçümü (F-measure), doğruluk (accuracy), dengeli doğruluk (balanced accuracy) ve ROC eğrisi altında kalan alan (AUC) kullanılmış olup bu ölçütler karışıklık matrisi kullanılarak hesaplanmaktadır.

Karışıklık Matrisi (Confusion Matrix): Karışıklık matrisi, etiketli verilerin sınıflandırılması sonucunda verilerin öngörülen sınıflarını ve gerçek sınıflarını içerir. Tablo 1'de iki sınıfa ait karışıklık matrisi sunulmuştur.

Tablo 1. İki sınıf için oluşturulmuş örnek bir karışıklık matrisi.

Table 1. An example confusion matrix created for two classes.

Karışıklık Matrisi		Gerçek Sınıf		
		Pozitif (Hasta)	Negatif (Sağlıklı)	Toplam
Tahmini Sınıf	Pozitif (Hasta)	A (DP)	B (YP)	A + B
	Negatif (Sağlıklı)	C (YN)	D (DN)	C + D
	Toplam	A + C	B + D	A+B+C+D

Bu tabloda yer alan DP (Doğru Pozitif) gerçekte hasta olup sınıflandırma sonucunda hasta olarak etiketlenen örnek sayısını gösterirken YN (Yanlış Negatif) gerçekte hasta olup sağlıklı olarak etiketlenmiş örnek sayısını göstermektedir. DN (Doğru Negatif), gerçekte sağlıklı olup sınıflandırma sonucunda sağlıklı olarak etiketlenen örnek sayısını gösterirken YP (Yanlış Pozitif) gerçekte sağlıklı olup hasta olarak etiketlenen örnek sayısını göstermektedir.

Duyarlılık (Sensitivity/Recall): Testin, gerçek hastalar içinden hastaları ayırma yeteneğidir [31].

$$\text{Duyarlılık} = \text{DP}/(\text{DP} + \text{YN}) \quad (1)$$

Seçicilik (Specificity): Testin, gerçek sağlamlar içinden sağlamları ayırma yeteneğidir [31].

$$\text{Seçicilik} = \text{DN}/(\text{DN}+\text{YP}) \quad (2)$$

F-Ölçümü (F-Measure): Bu ölçüt, kesinlik ve duyarlılık ölçütlerinin harmonik ortalamasıdır. Literatürde sıklıkla kullanılmaktadır. Çünkü sistemin başarısı değerlendirilirken tek başına kesinlik veya tek başına duyarlılık değerlendirmesi eksik kalmaktadır [31].

$$\text{F-Ölçümü} = 2^* (\text{Duyarlılık}^* \text{Kesinlik})/(\text{Duyarlılık} + \text{Kesinlik}) \quad (3)$$

Doğruluk (Accuracy): Duyarlılık ve seçicilik birleştirilerek tek bir ölçü elde edilmek istendiğinde kullanılan ölçülerden birisi de doğru test sonucu olasılığıdır. Gerçekte testin hasta ve sağlam olarak toplam doğru tanı koyma oranına denir [31].

$$\text{Doğruluk} = (\text{DP}+\text{DN})/(\text{DP}+\text{YP}+\text{DN}+\text{YN}) \quad (4)$$

Dengeli Doğruluk (Balanced Accuracy): Literatürde sıklıkla geleneksel doğruluk ölçütü kullanılmasına karşın dengeli doğruluk ölçütünün kullanılması daha tarafsız bir yaklaşımdır. Çünkü geleneksel doğruluk ölçütünde sadece tek bir sınıfa (hasta veya sağlıklı) ait doğruluk değeri ele alınırken dengeli doğrulukta her iki sınıftan (hasta ve sağlıklı) elde edilen doğruluğun ortalaması ele alınmaktadır. Yani hastaların doğru tahmin edilmesinin yanında sağlıklıların da doğru tahmin edilmesi gerekmektedir. Sınıflandırıcı her iki sınıfta da eşit derecede iyi performans gösteriyorsa, bu terim geleneksel doğruluk değerine göre daha düşük bir orana sahip olur [32].

$$\begin{aligned} Dengeli\ Doğruluk &= \frac{1}{2}\left(\frac{DP}{DP+YN} + \frac{DN}{DN+YP}\right) \\ &= \frac{1}{2}(Duyarlılık + Seçicilik) \end{aligned} \quad (5)$$

ROC eğrisi altında kalan alan (Area under curve-AUC): AUC, ROC [30] eğrisini özetleyen ortalama bir performans değeri verir. ROC eğrisinin altında kalan alan, testin hastalar ile hasta olmayan bireyleri ayırmadaki doğruluk oranını belirler. Bir sınıflandırıcı, ROC eğrisi sol üste ne kadar yakınsa, yani AUC değeri bire ne kadar yakınsa o kadar tercih edilir [33].

2.3 Materyal

Bu çalışmada kullanılan veri seti Kars ilinde bulunan iki koyun çiftliğinden elde edilmiştir. 301 Akkaraman cinsi koyun ve bunlardan doğan 347 kuzuya kulak küpesi uygulanarak hastalık ve diğer bilgiler arşivlenmiştir. Kan örnekleri doğumdan sonraki 24 sa. içinde alınmıştır. Kuzularda sağlık durumları yaşamın ilk 1 ayında (neonatal periyot) günlük, sonraki 8 haftalık dönemde (post-neonatal periyot) 2 günde bir yapılan ziyaretlerle ilgili araştırmacılar tarafından kayıt altına alınmıştır [34]. Çalışma süresi boyunca hastalık (mastitis, pneumonia, enteritis, pregnancy toxaemia etc.) olduğu tespit edilen koyunlar ve hastalık (ishal, pnömoni, septisemi, halsizlik, pmöoenteritis, vd.) olduğu tespit edilen kuzular hasta olarak etiketlenmiş ve kulak etiketi numarası ile kaydedilmiştir.

Kuzularda plesental yapıdan dolayı anneden yavruya başta hastalıklara karşı koruyucu antikorlar olmak üzere yaşam için gerekli olan birçok maddenin geçişi olmaz. Kuzuların hastalıklarından korunması ve normal gelişmesi için gerekli olan tüm maddeler annelerin doğumdan sonra ürettikleri ilk süt/kolostrumda bulunmaktadır. Bu nedenle kolostrumun yeterli alınması oldukça önemlidir ve yetersizliği doğumdan sonra ilk 24 sa. ölçülen IgG gibi çeşitli kan parametreleri (Tablo 2'de sunulmuştur) ile belirlenebilmektedir. Özellikle neonatal dönemde gelişen hastalıklarda kolostrumda bulunan maddelerin yeterli alınmasıyla doğrudan ilişkilidir. Ancak bu durum post-neonatal dönemde etkisini kaybederek hastalıkların gelişmesinde işletmenin fiziksel ve çevresel koşulları, aşılama gibi faktörlerin etkili olduğu bilinmektedir [35]. Bu nedenle çalışmada hastalık sınıflandırılmasının neonatal kuzular üzerinde gerçekleştirilmesine karar verilmiştir. Bu kapsamda analizlerin doğru bir şekilde yapılabilmesi için veri setinden neonatal kuzuların hastalık durumu ile ilişkisiz veya doğrudan hastalıklarla ilişkisi olan özellikler çıkartılmıştır. Sonuç olarak çalışmada 347 (60 hasta, 287 sağlıklı) örnek, 14 özellik ve 1 sınıf etiket bilgisi kullanılmıştır. Bu özelliklerin bilgisi Tablo 2'de sunulmuştur.

Tablo 2. Veri seti özeti [31],[36].

Table 2. Data set summary [31],[36].

Nümerik Değişkenler					
Özellikler	En Küçük	En Büyük	Ortalama	Ortalama (Hasta, n=60)	Ortalama (Sağlıklı, n=287)
IgG (Immunoglobulin G)****	19	5302	2196	1526±1268	3114 ± 1121
GGT (Gamma Glutamil Trasnfereraz)**	38	7517	2382	1780±1655	2855±1410
LT (Laktoferrin)**	354	2194	1052	955±312	1064±321
TP (Total Protein)****	21	117	73	62±18	78±11
ALB (Albumin)*	32	51	40	41±4	41±4
BW (Doğum ağırlığı)****	2260	5900	4028	3641±708	4143±593
WG28 (28. gün sonundaki vücut ağırlığı)****	4364	14016	8804	7377±1975	9378±1913
MDG28 (Ortalama günlük kilo kazanımı)****	14	340	169	130±53	185±55
Age (Anne yaşı)	1	6	3	3±1	3±1
Parity (Anne doğum sayısı)	1	5	2	2±1	2±1

Kategorik Değişkenler				
Özellikler	Hasta Kuzu (n = 60)		Sağlıklı Kuzu (n = 287)	
AH (Anne sağlık durumu)**	Hasta n=11	Sağlıklı n=49	Hasta n=6	Sağlıklı n=281
Twin (Doğum tipi)	İkiz n=19	Tek n=41	İkiz n=73	Tek n=214
Gender (Cinsiyet)	Erkek n=22	Dişi n=38	Erkek n=134	Dişi n=153
Farm (Çiftlik)*	Farm1 n=12	Farm2 n=48	Farm1 n=18	Farm2 n=269

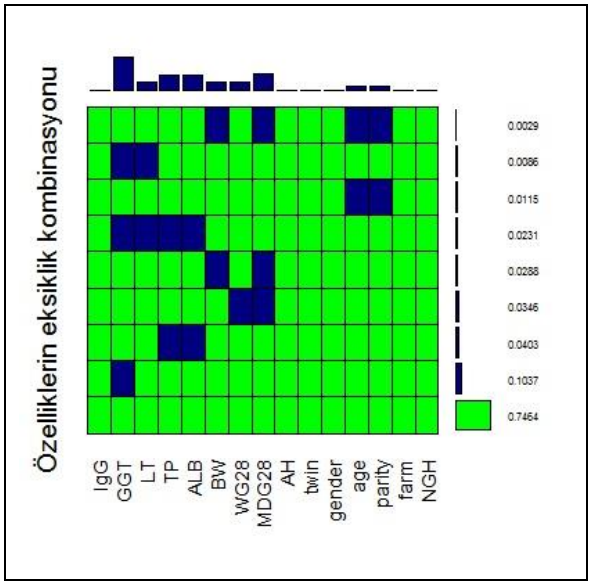
****: P<0.0001, ***: P<0.001, **: P<0.01, *: P<0.05, n=Hayvan sayısı.

3 Bulgular

Bilimsel araştırmalar kapsamında üzerinde çalışılmak istenilen veriler hastalık, ölüm, analizin hatalı yapılması, ölçümü yapılan örneğin uygun olmaması gibi nedenlerle istenildiği gibi eksiksiz bir şekilde toplanılamayabilir. Eksik veriler hemen hemen tüm araştırmaların bir parçası olup veri eksikliği, veterinerlik alanındaki çalışmalarda da sıklıkla görülmektedir.

Veri setlerindeki bu eksik değerler, kayıp değerler olarak adlandırılır ve bu durum birçok araştırmacının karşılaştığı bir dezavantajdır. Çünkü çoğu istatistiksel veri analizi paket programları verilerin kayıpsız olduğu varsayımı altında geliştirilmiştir.

Çalışmada kullanılan veri setindeki bazı özellikler eksik değerler içermektedir. Bu özelliklerdeki eksik değerlerin oranları ve birlikte ne oranda eksik veri içerdiği Şekil 2'de sunulmuştur.



Şekil 2. Özelliklerin eksik veri oranları ve kombinasyonları.

Figure 2. Missing data rates and combinations of variables.

Şekil 2'de sütunlar özelliklerin eksiklik oranlarını gösterirken satırlar özelliklerin birlikte eksik değer içerme oranını göstermektedir. Şekildeki sütunlar incelendiğinde IgG, anne hastalık durumu (AH), ikizlik (twin), cinsiyet (gender), çiftlik (farm) ve sınıf etiketi yani neonatal kuzunun hastalık durumunu gösteren özelliklerde tüm hücreler yeşil olup eksik değer içermemektedir. En fazla eksik değere GGT özelliği sahip olup onu sırasıyla MDG28, TP, ALB, WG28, LT, BW, yaş ve doğum sayısı özellikleri takip etmektedir. Eksik değer içeren özellikler içerisinde kan seviyelerini gösteren 5 özellikten (IgG, GGT, TP LT ve ALB) 4'nün eksik değer içerdiği ve GGT kan seviyesi özelliğinin diğer kan seviyelerine göre neredeyse iki katından fazla eksik değer içerdiği görülmektedir.

Şekil 2'deki satırlar incelendiğinde veri setindeki örneklerin yaklaşık %75'inin eksik değer içermediği, yaklaşık %10'nunda sadece GGT özelliğinin eksik olduğu, yaklaşık %3'ünde dört kan seviyesinin (GGT, TP, LT ve ALB) birlikte eksik olduğu, yaklaşık %3'ünde sadece yaş özelliğinin eksik olduğu görülmektedir. Sonuç olarak GGT özelliğindeki eksik değer oranı diğer özelliklerdeki eksik değer oranından 2 ile 10 kat daha fazla olduğu gözlemlenmiştir.

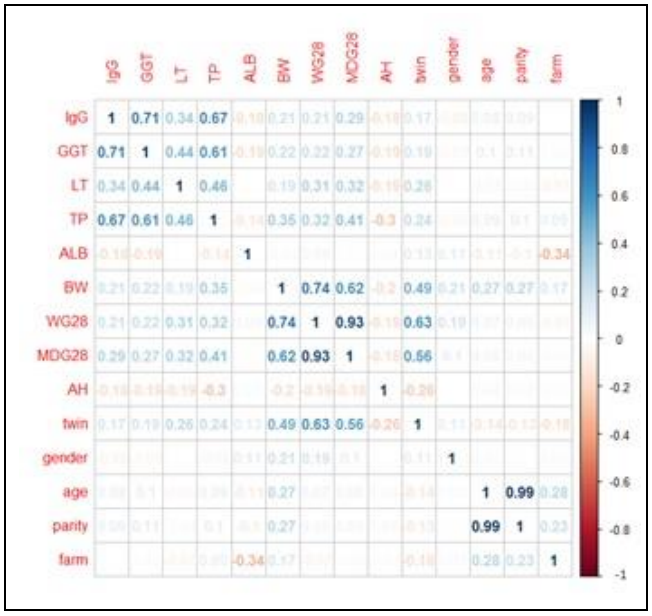
Günümüzde veri setindeki eksik değerleri tamamlamak için birçok geleneksel ve modern eksik veri tamamlama yöntemleri kullanılmaktadır. Bu çalışmada kullanılan veri setindeki eksik değerler yapay arı kolonisi (ABC) yöntemi ile tamamlanmıştır [31]. Veri setindeki eksik değerler tamamlandıktan sonra veri setinin özeti Tablo 2'de sunulmuştur. Çalışmada bağımsız grup t-testi (Independent Samples t Test) kullanılmıştır.

Öncelikle, kuzularda hastalık risk faktörlerini en basit düzeyde inceleyecek olursak; hasta olan grup ile sağlıklı olan grubun IgG, GGT, LT, ALB, TP, BW, WG28, MDG8, AH ve farm değerleri ortalamaları istatistiksel olarak anlamlı farklılık göstermiştir. Risk faktörleri ile hastalık varlığının ilişkisi incelendiğinde;

- Hastalık varlığı ile IgG seviyesinin düşük olması arasında anlamlı ilişki saptanmıştır. Hasta olanların ortalama IgG seviyesi 1526±1268 iken, sağlıklı olan kuzuların ortalama IgG seviyesi 3114 ± 1121 olarak saptanmıştır (P=1.768e-05),

- Hastalık varlığı ile GGT seviyesinin düşük olması arasında anlamlı ilişki saptanmıştır. Hasta olanların ortalama GGT seviyesi 1780±1655 iken, sağlıklı olan kuzuların ortalama GGT seviyesi 2855±1410 olarak saptanmıştır (P=0.002),
- Aynı şekilde LT (P=0.009), TP (P=2.753e-07), ALB (P=0.0397), BW (P=1.086e-05), WG28 (P=02.514e-08), MDG28(P=1.014e-08), özellikleri de istatistiksel olarak anlamlı bulunmuştur,
- Diğer risk faktörlerinden anne yaşı, anne doğum dayısı, ikizlik ve cinsiyet ile hastalık varlığı arasında istatistiksel olarak anlamlı ilişki saptanmamıştır.

Veri setindeki özelliklerin ilişki matrisi Şekil 3'te gösterilmiştir.



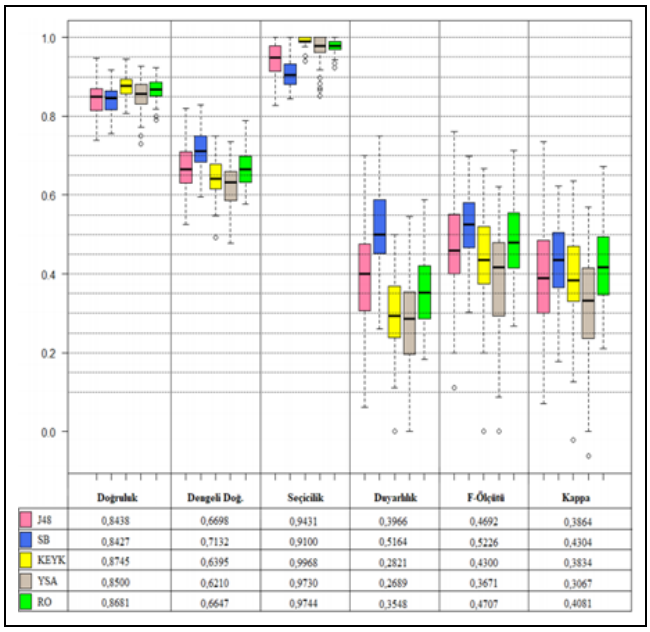
Şekil 3. Değişkenler arasındaki ilişki matrisi [31].

Figure 3. Relationship matrix between variables [31].

Korelasyon matrisindeki koyu mavi renk değişkenler arasındaki güçlü pozitif ilişkiyi gösterirken koyu kırmızı renk ise güçlü negatif ilişki olduğunu göstermektedir. Değişkenler arasındaki ilişkinin gücü arttıkça renk koyulaşmaktadır. Korelasyon matrisi incelendiğinde IgG özelliğinin hem GGT hem de TP özelliği ile arasında yüksek korelasyon olduğu görülmektedir.

Neonatal kuzular hastalık durumuna göre sınıflandırıldığında sınıflandırma yöntemlerinin doğruluk, dengeli doğruluk, seçicilik, duyarlılık, F-Ölçütü ve kappa sonuçları Şekil 4'te sunulmuştur.

Birçok çalışmada sadece doğruluk ölçütü değeri göz önüne alınıp, dengeli doğruluk değeri dikkate alınmadan en yüksek doğruluk değerine sahip olan model en iyi performansa sahip model olarak nitelendirilir [19],[37]. Oysa sınıflar arası uygun dağılım gösteren dengeli veri setinde doğruluk değeri, dengesiz veri setine göre daha iyi performans göstermektedir. Eşit sınıf dağılımına sahip veri setinde doğruluk değeri ile dengeli doğruluk değeri aynı sonuca sahip olurken, veri setinin dengesiz olması durumunda ise dengeli doğruluk değeri doğruluk değerinden daha düşüktür. Nitekim neonatal kuzularda hastalık sınıflandırması sonuçları incelendiğinde (Şekil 4) en yüksek doğruluk değerine göre en iyi performansı KEYK ve RO modelleri gösterirken, dengeli doğruluk kriterine göre bu durum değişerek en iyi performansı SB modelinin gösterdiği görülmektedir.



Şekil 4. Sınıflandırma yöntemlerinin performans sonuçları [31].

Figure 4. Performance results of classification methods [31].

Seçicilik, gerçekte sağlıklı olan kuzuların sistem tarafından da hasta olarak tahmin edildiği durumdur. KEYK yönteminin sağlıklı kuzuları %99.7 başarı oranı ile tespit ettiği görülmektedir. Her ne kadar sağlıklı kuzuları doğru tahmin etmek önemli olsa da bizim için asıl önemli olan hastalanabilecek kuzuların tespit edilerek önlemler alınmasıdır. Bu nedenle duyarlılık ölçüsü önemsenecek ve göz önünde bulundurulacak kıstastır.

Duyarlılık, gerçekte hasta olan kuzuların sistem tarafından hasta olarak tahmin edildiği durumdur. SB yönteminin hasta kuzuları %52 başarı oranı ile tespit ettiği görülmektedir. Veri setindeki 347 kuzudan 60 tanesi hastadır. Kuzuların içinden hastayı tahmin etme ihtimali normalde yaklaşık %17 iken, veri madenciliği yöntemi ile bu oran yaklaşık %52'dir. Hasta olabilecek kuzuların önceden tahmin edilmesi, erken tedavi için büyük öneme sahiptir.

F-ölçütü, kesinlik ve duyarlılığın ağırlıklandırılmış ortalamasını göstermektedir. Başlı başına kesinlik ve duyarlılığa bakmak yerine F-ölçütüne bakmak daha doğru sonuç vermektedir. %52 F-ölçütü oranıyla en yüksek başarıyı yine SB yöntemi göstermiştir.

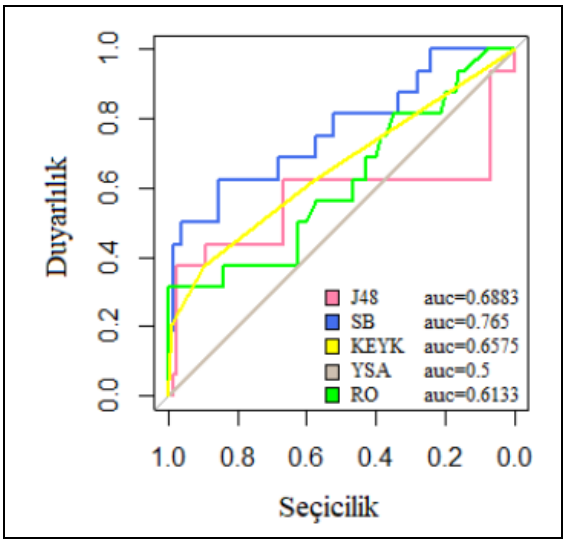
Kappa değeri sınıflandırıcı modelin doğru sınıflandırma başarımı hakkında bilgi vererek başarının şans faktörüne bağlı olup olmadığı hakkında fikir vermektedir. SB yöntemi yaklaşık 0.435 kappa değerine sahiptir. Kappa değerinin 0.40'dan büyük olmasının makul olduğu araştırmacılar tarafından bildirilmiştir [38]. SB modelinin 0.435 kappa değeri nedeniyle tutarlı tahminler yaptığını göstermektedir.

Özellikle sağlık alanında yapılan modellerin değerlendirilmesinde sıklıkla kullanılan AUC kriteri sonuçları Şekil 5'te gösterilmiştir.

Sınıflandırma yöntemlerinin AUC grafiği incelendiğinde, 0.765 oranı ile SB yönteminin diğer yöntemlerden daha başarılı

olduğu görülmektedir. Buda bize SB modelinin verileri sınıflandırmada ayrım gücünün yüksek olduğunu gösterir.

Çalışmada kullanılan veri setindeki özellik sayısı ve örnek sayısı azdır. Büyük veri setlerinde kernel yöntemler başarılı sonuçlar vermektedir. Klasik sınıflandırıcılar model öğrenirken hatayı minimize etmektedirler ve parametrelerin optimizasyonunda baskın sınıfı teşvik edecek bir yanlılığa sebep almaktadır. SB yönteminde, optimize edecek bir üst parametre olmamakla beraber olasılık dağılımlarının her hedef değişken için ayrı ayrı ve sonsal olasılığın sınıf önsel olasılıklarıyla ağırlıklandırılarak hesaplanması, azınlık sınıfların tanınması için diğer yöntemlere göre daha fazla şans vermektedir.



Şekil 5. Sınıflandırma yöntemlerinin AUC sonuçları [31].

Figure 5. AUC results of classification methods [31].

4 Sonuçlar

Neonatal kuzularda hastalık sınıflandırması için; karar ağacı (J48), Saf bayes, k-en yakın komşu, yapay sinir ağı ve rastgele orman sınıflandırma algoritması kullanılmıştır. Sınıflandırıcıların performansları doğruluk, dengeli doğruluk, seçicilik, duyarlılık, f-ölçütü, kappa ve AUC ölçütlerine göre karşılaştırılmıştır. Model başarım ölçütleri incelendiğinde, %84 doğruluk, %71 dengeli doğruluk, %52 duyarlılık, %52 f-ölçütü, 0.4304 kappa ve 0.765 AUC değeri ile en başarılı yöntemin Saf Bayes olduğu gözlemlenmiştir. Kappa değerinin 0,40'dan büyük olması modelin tutarlı tahminler yaptığını göstermektedir. Özellikle sağlık alanında yapılan modellerin değerlendirilmesinde kullanılan AUC değerinin 0,765 olması geliştirilen modelin verileri sınıflandırmada ayrım gücünün yüksek olduğunu belirtir. Veri setinde 287 sağlıklı ve 60 hasta kuzu olmasından dolayı dengeli bir sınıf dağılımı söz konusu değildir. Nitekim doğruluk değerinin, dengeli doğruluk değerinden daha yüksek olması bize dengesiz bir sınıf dağılımının olduğunu göstermektedir. 347 (60 hasta, 287 sağlıklı) kuzu içerisinden hasta kuzuları tahmin etme oranı yaklaşık %17 iken veri madenciliği yöntemi ile bu oran yaklaşık %52'dir. Bilgisayar destekli tanı ile neonatal kuzulardan hasta olabilecekler tahmin edilerek erken teşhis ve tedavide veteriner hekime yardımcı olunabilecektir. Erken teşhis ve tedavi sayesinde hastalıklar ve ölümlerdeki azalma ile ülke ekonomisine katkı sağlanabilecektir.

5 Teşekkür

Bu çalışmada kullanılan veri seti TÜBİTAK projesi (Proje Kodu: TOVAG 108 O 847) kapsamında toplanmıştır.

6 Kaynaklar

[1] Esfandiari N, Babavalian MR, Moghadam A-ME, Tabar VK. "Knowledge discovery in medicine: Current issue and future trend". *Expert Systems with Applications*, 41(9), 4434-4463, 2014.

[2] Bennett TD, Callahan TJ, Feinstein JA, Ghosh D, Lakhani SA, Spaeder MC, Kahn MG. "Data science for child health". *The Journal of Pediatrics*", 208,12-22, 2019.

[3] Miller DD, Brown EW. "Artificial intelligence in medical practice: the question to the answer?". *The American Journal of Medicine*, 131(2), 129-133, 2018.

[4] Gülten A, Doğan Ş. "Genetik algoritmalar yönteminin biyomedikal verileri üzerindeki uygulamaları". *Doğu Anadolu Araştırma ve Uygulama Merkezi Dergisi*, 7(1), 12-16, 2009.

[5] Haupt RL, Haupt SE. *Practical Genetic Algorithms*. 2nd ed. New York, USA, Wiley, 2004.

[6] Han J, Pei J, Kamber M. *Data Mining: Concepts and Techniques*. 3rd ed. Amsterdam, Elsevier, 2011.

[7] Tan PN. *Introduction to Data Mining*. India, Pearson Education, 2006.

[8] Cihan P, Gökçe E, Kalıpsız O. "A review of machine learning applications in veterinary field". *Kafkas University Veterinary Faculty Journal.* 23(4), 673-680, 2017.

[9] Zarchi HA, Jonsson R, Blanke M. "Improving oestrus detection in dairy cows by combining statistical detection with fuzzy logic classification". *Proceedings Workshop on Advanced Control and Diagnosis*, Zielona Gora, PL, 1-20 November 2009.

[10] Memmedova N, Keskin İ. "Oestrus detection by fuzzy logic model using trait activity in cows". *Kafkas University Veterinary Faculty Journal*, 17(6), 1003-1008, 2011.

[11] Brown-Brandl T, Jones DD, Woldt W. "Evaluating modelling techniques for cattle heat stress prediction". *Biosystems Engineering*, 91(4), 513-524, 2005.

[12] Dórea JRR, Rosa GJM, Weld KA, Armentano LE. "Mining data from milk infrared spectroscopy to improve feed intake predictions in lactating dairy cows". *Journal of Dairy Science*, 101(7), 5878-5889, 2018.

[13] A, Atıl H, Kesenkaş H. "Çiğ süt kalite değerlendirmesinde bulanık mantık yaklaşımı". *Kafkas Universitesi Veteriner Fakultesi Dergisi*, 20(2), 223-229, 2014.

[14] Mehraban Sangatash M, Mohebbi M, Shahidi F, Vahidian Kamyad A, Qhods Rohani M. "Application of fuzzy logic to classify raw milk based on qualitative properties". *International journal of AgriScience*, 2(12), 1168-1178, 2012.

[15] Akıllı A, Atıl H, Kesenkaş H. "Çiğ süt kalite değerlendirmesinde bulanık mantık yaklaşımı". *Kafkas Üniversitesi Veteriner Fakültesi Dergisi*, 20(2), 223-229, 2014.

[16] Küçükönder H, Üçkardeş F, Narinç D. "Hayvancılık alanında bir veri madenciliği uygulaması: Japon bıldırcını yumurtalarında döllülüğe etki eden bazı faktörlerin belirlenmesi". *Kafkas Universitesi Veteriner Fakultesi Dergisi*, 20(6), 903-908.

[17] Kılıç İ, Özbeyaz C. "Bulanık kümeleme analizinin koyun yetiştiriciliğinde kullanımı ve bir uygulama". *Kocatepe Veteriner Dergisi*, 3(2), 31-37, 2010.

[18] Lyytikäinen T, Kallio E. "Risk classification of Finnish pig farms by Simulated FMD Spread". *In Proc. Society of Veterinary Epidemiology and Preventive Medicine annual meeting*, Liverpool, UK, 25-28 March 2008.

[19] Awaysheh A, Wilcke J, Elvinger F, Rees L, Fan W, Zimmerman KL. "Evaluation of supervised machine-learning algorithms to distinguish between inflammatory bowel disease and alimentary lymphoma in cats". *Journal of Veterinary Diagnostic Investigation*, 28(6), 679-687, 2016.

[20] Dobbin KK, Simon RM. "Optimally splitting cases for training and testing high dimensional classifiers". *BMC Medical Genomics*, 4(31), 2-8, 2011.

[21] Dietterich TG. "Approximate statistical tests for comparing supervised classification learning algorithms". *Neural Computation*, 10(7), 1895-1923, 1998.

[22] Team RC. "R: A Language and environment for statistical computing". R Foundation for Statistical Computing, Vienna, Austria, 2013.

[23] Safavian SR, Landgrebe D. "A survey of decision tree classifier methodology". *IEEE Transactions on Systems, Man, and Cybernetics*, 21(3), 660-674, 1991.

[24] Aydoğan Ü. Destek Vektör Makinalarında Kullanılan Çekirdek Fonksiyonların Sınıflama Performanslarının Karşılaştırılması. Yüksek Lisans Tezi, Hacettepe Üniversitesi, Ankara, Türkiye, 2010.

[25] Xiang S, Nie F, Zhang C. "Learning a Mahalanobis distance metric for data clustering and classification". *Pattern Recognition*, 41(12), 3600-3612, 2008.

[26] Wang H. "Nearest neighbors by neighborhood counting". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6), 942-953, 2006.

[27] Garcia V, Debreuve E, Barlaud M. "Fast k nearest neighbor search using GPU". *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops,* Anchorage, Alaska, USA, 23-28 Jun 2008.

[28] Kolahdouzan MR, Shahabi C. "Continuous k-nearest neighbor queries in spatial network databases". *Spatio-Temporal Database Management, 2nd International Workshop STDBM'04*, Toronto, Canada, 30 August 2004.

[29] Alpaydin E. *Introduction to Machine Learning*, 2nd ed. Cambridge, Massachusetts, London, England, MIT Press, 2010.

[30] Breiman L. "Random forests". *Machine Learning*, 45(1), 5-32, 2001.

[31] Cihan P. Determination of diagnosis, prognosis and risk factors in animal diseases using by data mining methods. PhD Thesis, Yildiz Technical University, Istanbul, Turkey, 2018.

[32] Brodersen KH, Ong CS, Stephan KE, Buhmann JM. "The balanced accuracy and its posterior distribution". *Pattern recognition (ICPR)*, 3121-3124. IEEE, 2010.

[33] DeLong ER, DeLong DM, Clarke-Pearson DL. "Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach". *Biometrics,* 44(3), 837-845, 1988.

[34] Gökçe E, Erdoğan HM. "An epidemiological study on neonatal lamb health". *Kafkas Üniversitesi Veteriner Fakültesi Dergisi*, 15(2), 225-236, 2009.

[35] Gökçe E, Kırmızıgül AH, Erdoğan HM, Citil M. "Risk factors associated with passive immunity, health, birth weight and growth performance in lambs: I. effect of parity, dam's health, birth weight, gender, type of birth and lambing season on morbidity and Mortality". *Kafkas Üniversitesi Veteriner Fakültesi Dergisi*, 19(Suppl-A), A153-A160, 2013.

[36] Cihan P, Kalıpsız O, Gökçe E. "Hayvan hastalığı teşhisinde normalizasyon tekniklerinin yapay sinir ağı ve özellik seçim performansına etkisi". *Electronic Turkish Studies*, 12(11), 2017.

[37] Brewster LR, Dale JJ, Guttridge TL, Gruber SH, Hansell AC, Elliott M, Gleiss AC. "Development and application of a machine learning algorithm for classification of elasmobranch behaviour from accelerometry data". *Marine Biology*, 165(4), 62, 2018.

[38] Fleiss JL, Levin B, Paik MC. *Statistical Methods for Rates and Proportions.* 3rd ed. New York, USA, John Wiley & Sons, 2013.