Regular paper | Journal of the KIECS. pp. 1157-1170, vol. 18, no. 6, Dec. 31. 2023, t. 122, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2023.18.6.1157

랜덤 포레스트 기계 학습 방법을 이용한 넙치의 복수 증상 분석

김경임* · 김성현** · 정희택*** · 한순희*** · 박정선***†

Analysis of Ascites Symptoms in Cultured Olive Flounder, *Paralichthys Olivaceus,* using a Random Forest Machine Learning Method

Kyeong-Im Kim* · Sung-Hyun Kim** · Hee-Taek Ceong*** · Soonhee Han*** · Jeong-Seon Park***†

요 약

복수는 물고기의 복강에 체액이 비정상적으로 축적되는 상태로써 넙치의 건강 상태를 나타내는 중요한 지표이다. 박테리아, 바이러스, 기생충 등의 감염 과정에서 복수가 생길 수 있으며, 이로 인해 복부의 팽만, 부진한 성장 및 체중 감소 등이 나타난다. 본 논문에서는 넙치의 복수 증상에 영향을 미치는 다른 증상 또는 질병과의 연관성을 찾고자 하였다. 실험 데이터로는 복수의 증상을 복수 없음, 복수 투명, 복수 불투명의 3가지 상태로 구분하고 7년 동안 수집한 양식넙치의 질병진단 데이터를 사용하였다. 랜덤 포레스트 기계 학습 방법을 위해 적절한 전처리 과정을 수행한 후 복수 증상과 관련 있는 다른 증상 및 질병 인자들을 추출하였으며, 제안된 모델이 복수 증상 관련 주요 인자들을 제시해 줄 수 있음을 확인하였다.

ABSTRACT

Ascites is a condition in which body fluids are abnormally accumulated in the fish's abdominal cavity, and is an important indicator of the health of flounder. Ascites can occur in the process of infection with bacteria, viruses, parasites, etc., which causes abdominal distension, sluggish growth, and weight loss. In this paper, we tried to find the correlation with other symptoms or diseases that affect ascites symptoms in flounder. As experimental data, ascites symptoms were divided into three states: no ascites, ascites transparent, and ascites opaque, and disease diagnosis data of cultured flounder collected for 7 years were used. After performing an appropriate preprocessing process for the random forest machine learning method, other symptoms and disease factors related to ascites were extracted, and it was confirmed that the proposed model could present the main factors related to ascites.

키워드

Paralichthys olivaceus, Clinical Signs of Ascites, Machine Learning, Random Forest
양식 넙치, 복수의 증상, 기계 학습, 랜덤 포레스트

* 스마트수산양식연구센터 연구원(insungup@hanmail.net)
** 수산질병관리진단전문연구소 피쉬케어 소장
(sunghyun.kim@live.co.kr)
*** 전남대학교 문화콘텐츠학부 교수
(htceong@chonnam.ac.kr, shhan@chonnam.ac.kr)
† 교신저자 : 전남대학교 문화콘텐츠학부 교수
· 접 수 일 : 2023. 08. 29
· 수정완료일 : 2023. 10. 20
· 게재확정일 : 2023. 12. 27
· Received : Aug. 29, 2023, Revised : Oct. 20, 2023, Accepted : Dec. 27, 2023
· Corresponding Author : Jeong-Seon Park
Division of Culture Contents, Chonnam National University,
Email : jpark@jnu.ac.kr

Ⅰ. 서 론

복수는 물고기의 복강에 체액이 비정상적으로 축적되는 상태로써 넙치의 건강 상태를 나타내는 중요한 지표이다. 양식업자와 수산질병관리사는 복수를 인식함으로써 넙치의 잠재적인 건강 문제를 식별하고 근본적인 원인을 진단하고 치료하기 위한 적절한 조치를 할 수 있다[1].

넙치에 복수 증상이 발생하면 복부의 팽만, 비정상적인 신체 자세, 활동 감소, 호흡 곤란, 부진한 성장 및 체중 감소 등의 변화가 나타난다. 또한 다양한 박테리아, 바이러스, 기생충 감염 등의 과정에서 복수 증상이 나타나기도 한다. 국립수산과학연구소의 보고서에 따르면 양식 넙치에서 복수 증상과 관련 있는 질병으로는 랍도바이러스병(Rhabdovirus olivaceus), 비루나바이러스병(Birnavirus), 연쇄구균증(Streptococcus sp.), 에드워드병(Edwardsielosis) 등이 있다.

본 연구에서는 양식 넙치의 질병을 진단하는 데 있어 중요한 복수의 증상과 다른 증상 또는 질병과의 연관성을 찾고자 하였다. 이를 위해 복수의 상태를 복수 없음, 복수 투명, 복수 불투명의 3가지 상태로 구분하고 2015년부터 7년 동안 정기적으로 수집한 양식넙치의 질병진단 데이터를 사용하였다[1, 2]. 또한 기계학습 분야에서 예측 모델링 및 데이터 분석에 강력하고 유연하다고 알려진[3, 4] 랜덤 포레스트(Random Forest) 모델을 이용하여 복수 증상과 관련 있는 다른 증상 및 질병 인자들을 추출하고, 분류 실험을 통해 제안된 모델이 복수 증상 관련 주요 인자들을 제시해 줄 수 있음을 확인하고자 하였다.

본 논문의 구성은 다음과 같다. 2장에서는 관련 연구로서 복수 증상 분석의 중요성과 기계학습을 이용한 질병 진단 사례, 그리고 랜덤 포레스트 모델을 이용한 사례를 알아보고, 3장에서는 복수 증상 분석을 위한 실험 데이터에 대해 자세히 알아본다. 4장에서는 파이썬 라이브러리를 활용하여 랜덤 포레스트 모델을 이용한 복수 증상 분석 실험을 수행한 결과를 분석하고, 마지막으로 5장에서 결론 및 향후 연구 방향에 대해 논의한다.

1) https://nifs.go.kr/fishguard/disease02

Ⅱ. 관련 연구

2.1 복수 증상 분석의 중요성

복수는 박테리아 또는 바이러스 감염, 간 또는 신장 기능 장애, 심혈관 문제 또는 심지어 기생충 감염과 같은 다양한 근본적인 질병의 징후일 가능성이 있다. 그림 1은 피쉬케어연구소에서 수집한 질병진단 사례로 에드워드병에 감염되어 복수와 탈장 증상을 보이는 양식넙치의 사진을 보여준다[2].

그림 1. 에드워드균에 감염되어 탈장과 복수 증상을 보이는 양식 넙치의 예
Fig. 1 Examples of olive flounder infected with Edwardsiella piscicida showing hernia and ascites symptoms

또한, 복수는 종종 넙치의 사망률 증가와 관련이 있다. 이 상태는 생리적 불균형, 손상된 장기 기능 및 항상성 유지 능력 감소로 이어질 수 있으며, 치료하지 않고 방치하면 영향을 받은 넙치는 호흡 곤란, 성장 감소, 섭식 장애를 경험하고 결국 사망에 이르게 된다. 따라서 복수 증상의 조기 발견 및 조치는 양식업에서 사망 위험을 완화하고 경제적 손실을 최소화하는 데 도움이 될 수 있다.

복수는 넙치의 감염성 질병의 지표 역할을 할 수 있다. 박테리아나 바이러스와 같은 특정 병원균은 복강에 염증과 체액 축적을 일으킬 수 있다. 양식업자는 복수 사례를 모니터링하고 적절한 진단 테스트를 수행함으로써 잠재적인 질병 발병을 식별하고 표적 질병 통제 조치를 시행할 수 있다. 질병의 조기 발견과 즉각적인 개입은 병원균의 확산을 최소화하

고 어류 개체군의 전반적인 건강을 유지하는 데 매우 중요하다.

요약하면, 넙치의 복수 존재를 분석하는 것은 어류 건강 평가, 사망 위험 관리 및 질병 감시 수행에 필수적이다. 복수 증상을 인식하고 근본적인 원인을 해결함으로써 양식업자는 양식장의 운영 과정에서 넙치의 건강과 생산성을 향상할 수 있다.

2.2 기계 학습을 이용한 어류 질병 분석 사례

기계 학습을 이용한 어류 질병 검출 관련 연구의 대부분은 영상 처리 기법을 사용하였다. 초기의 연구로는 V. Lyubchenko 등이 색상 분할을 이용해서 어류의 표면에 보이는 백점(Ichthyophthirius multifiliis), 손상된 피부 및 누비병(quilted disease)을 검출하는 연구를 제안하였다[5]. 그러나, 초기에 사람이 개입하여 각각의 질병에 대한 흰색, 파란색, 빨간색 마커를 지정한 다음, 유사한 영역을 검출하고 어체의 감염 비율을 계산함으로써 질병 가능성을 예측하였으며, 실제 어류 영상이 아닌 인터넷 영상을 사용했다는 한계가 있다.

S. Malik 등은 Epizootic Ulcerative Syndrome(EUS)에 질병의 감염 여부를 판단하기 위하여 감염된 어류 DB의 이미지로부터 FAST(Features from Accelerated Segment Test) 특징을 추출하고 주성분 분석(Principal Component Analysis, PCA)으로 특징 차원을 축소한 다음 기계 학습 중의 하나인 신경망(Neural Network)을 사용하였으며, 분류 성능은 86.0%이라고 보고하였다[6].

유사한 연구 결과로 방글라데시의 연구팀은 다양한 영상 처리 기법과 기계 학습 방법의 하나인 SVM(Support Vector Machine)을 이용해 질병에 감염된 연어를 판별하는 사례가 있다. 이 연구도 자체 제작한 데이터에서 90% 이상의 분류 성능을 보인다고 발표하였다[7].

최근에는 딥러닝 기법을 통한 질병 예측 방법의 연구가 활발히 진행되고 있다. 손현승 등은 수산 양식장에서 발생할 수 있는 넙치의 질병을 딥러닝 기술로 예측하기 위해 양식장에서 수집된 카메라 영상에 데이터 증강과 전처리 포함하여 양식장의 수질 환경 및 생육 어류의 상태를 실시간 모니터링하는 연구 결과를 발표하였다[8]. 이 연구는 질병에 감염된 넙치의 영상을 인식하기 위해 YOLOv4 모델을 사용하여 딥러닝을 수행함으로써 질병의 발생 가능성을 예측하였다.

최근의 연구로는 Li 등이 딥러닝을 이용해 금붕어에서 발생하는 기생충 중에 Ichthyophthirius (Ichthyophthirius multifiliis), Monogenea (Gyrodactylus kobayashii), fish lice (Argulus japonicus)의 3종류의 기생충을 검출하고 기생충의 수를 세는 연구 결과를 발표하였다[9]. 이를 위해서 금붕어로부터 수집한 현미경 영상으로부터 fish lice, Monogenea 및 Ichthyophthirius를 포함하는 1,181개의 원본 이미지를 바탕으로 이미지 증강 등을 통해 생성한 기생충 이미지를 학습에 사용하였다. 또한 객체 검출을 위한 딥러닝 방법으로 YOLOv4를 이용하였으며, 자체 테스트를 통해 95% 이상을 정확도를 얻었다고 발표하였다.

지금까지 살펴본 대부분의 사례는 영상 데이터를 기반으로 질병 감염 여부를 판단하거나, 환경 정보를 포함하여 질병 가능성을 예측하거나 3종류 이내의 질병을 분류하는 연구이다.

본 연구팀은 이전 연구로써 실제 양식넙치의 외관, 해부, 현미경 검사 등을 통해 질병을 진단한 사례를 바탕으로 다양한 수온 구간에 따른 에드워드병의 증상 패턴을 분석하는 연구를 수행하였다[3]. 실제 진단 사례를 바탕으로 의사결정 나무 기법을 통해 진단 초기에 활용할 수 있는 에드워드병의 주요 증상을 찾아낸 연구이다.

2.3 랜덤 포레스트 관련 연구

랜덤 포레스트는 그림 2와 같이 주어진 데이터로부터 랜덤 샘플링을 통해 생성한 다수의 의사결정나무를 결합하여 최종 결과를 도출하는 기법으로 다수의 예측기 또는 분류기를 사용하여 정확도를 높이는 방법인 앙상블(ensemble)이 적용된 기계 학습 모델의 하나이다[3, 10]. 랜덤 포레스트는 의사결정나무 모델의 다양성을 증가시켜 과적합(overfitting)의 문제를 완화하고 예측 성능을 향상하는 장점이 있어 여러 분야에서 다양한 응용문제를 해결하기 위해 꾸준하게 사용되고 있다[11-13].

Islam 등은 특정 수중 환경에서 양식에 적합한 어종을 예측하기 위해 11종의 서로 다른 어류에 대한 수중 환경 데이터 세트를 구축하고 예측한 사례도

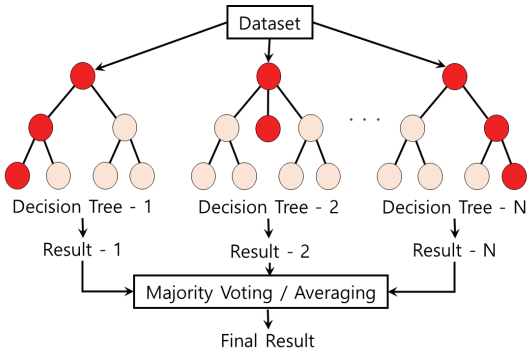

그림 2. 랜덤 포레스트 모델의 구조 예시
Fig. 2 An example of the structure of random forest model

있다[11]. 또한 Luan 등은 새우, 게 등의 수중 저서생물의 종 분포(풍부도, abundance)를 조사하는 데 있어서 샘플 크기에 따른 랜덤 포레스트 모델의 예측 능력을 측정하였다[12]. 10개부터 80개까지의 샘플 크기에 따른 성능 평가를 통해, 샘플 크기가 30개까지 증가할 때 성능의 증가 폭이 크며, 그 이후는 성능의 증가폭이 크지 않다고 보고했다. 또한 이주 행동, 수명, 신체 크기, 섭식 방식 및 유병률이 모델의 예측 성능에 영향을 미치는 주요한 요소라고 발표하였다.

스마트 수산 양식 분야에서도 랜덤 포레스트 모델이 사용된다. 최근의 연구에서는 어류 양식에서 성장에 영향을 미치는 중요한 요소인 염도, pH, 용존산소(DO), 온도와 같은 수질 매개변수의 시계열 데이터를 예측하기 위하여 랜덤 포레스트 모델을 사용하였으며, 2016년부터 2018년까지 24시간 간격으로 수집된 데이터를 이용해 모델을 구축하고 테스트하였다[13].

III. 실험 데이터

3.1 데이터 수집 및 전처리

양식 넙치의 복수 증상에 영향을 미치는 요소들을 분석하기 위해 피쉬케어연구소[2]에서 2015년부터 2021년까지 7년 동안 수집한 양식 넙치의 질병 진단 데이터를 사용하였다. 제주지역 39곳의 양식장을 주 1회 정기적으로 방문하여 질병을 모니터링하고 대량 폐사가 발생하지 않은 상황에서도 외관상 건강 상태가 상대적으로 나빠 보이는 개체를 1회당 5마리씩 채취하여 증상을 검사하고 질병을 진단한 데이터이다[2, 3].

수집한 양식 넙치의 질병 진단 데이터는 MS워드 파일로 저장된 데이터로 본 연구에 사용하는데 적합한 형태로 변환하였다. 먼저 파서 프로그램을 작성하여 의미 있는 데이터를 분리하고, 데이터 분석 분야에서 쉽게 처리할 수 있는 엑셀 문서로 저장하였다. 데이터 분석에 사용한 연도별 실험 데이터는 표 1과 같다.

표 1. 실험 데이터
Table 1. Experimental data

Year	Number of source data
2015	166
2016	899
2017	2,302
2018	3,687
2019	4,219
2020	4,046
2021	4,915
Total	20,234

실험 데이터는 양식 넙치의 질병 진단 결과를 총 252개의 세부 항목으로 분류하여 상세하게 기록한 자료이다. 이 중 양식장 이름, 날짜, 시간, 질병 기록, 수조 정보, 진단 소견 등은 텍스트 형태의 서술 항목이다. 측정값을 숫자로 기록한 항목은 수온, 길이, 무게 3개이고 어종 구분을 위해서는 (터봇 0, 넙치 1, 우럭 2, 다금바리 3, 도다리 4, 장어 5, 그루퍼 6, 돌돔 7, 강도다리 8) 정수를 사용한다. 10개를 제외한 242가지의 항목은 속성의 만족 여부를 각각 1과 0으로 표시하고 있다. 넙치의 질병과 관련된 증상은 외부, 내부, 현미경 관련, 병원체 등의 4개 그룹으로 분류한 후 최대 15개 조직과 관련된 세부 항목으로 나누어 정리하였다.

외부증상은 체표, 복부, 안구, 아가미, 지느러미 등의 15개 조직과 관련된 43개 증상(흑화, 발적, 팽

2) http://www.fishcare.kr/

창 등), 내부증상은 간, 장, 복수 등 11개 조직과 관련된 34개의 증상(출혈, 비대, 불투명 등)을 포함한다. 질병은 비브리오균(*Vibrio* spp.), 슈도모나스균(*Pseudomonas* spp.), 에드워드균(*Edwardsiella piscicida)*, 연쇄구균(*Streptococcus parauberis)*의 4개의 병원체에 대한 다량, 소량, 중량으로 구분하는 12개의 항목으로 구성되어 있다. 현미경 검경을 통해 검출된 증상(Micro related symptoms)에는 Micro surface 14개, Micro gill 13개, Micro brain 1개, Micro intestine 6개, Micro ascites 3개, Micro eye 2개, Micro ulcer 11개, Micro infection 1개가 있다. 이들은 다시 증상별로 다량, 소량, 중량의 153개로 나누어 기록하고 있다. 따라서 질병 관련 자료는 총 55개 증상 165가지 경우로 252개 항목의 65%를 차지한다.

3.2 복수 증상 데이터 분석을 위한 전처리

데이터 분석과 연관된 형태로 보면 전체적인 자료는 6개의 문자열 항목과 4개의 수치형 항목, 242개의 불린형으로 구성되어 있다. 2020년 이전에는 측정 항목에 대한 세분화가 부족하여 현재의 252개 항목 중 10개 항목에 대해 측정치가 없는 상태이다. 표 2는 2020년부터 추가된 10개의 증상 항목을 보여주는데, 5개는 외부증상이며, 나머지 5개는 내부증상이다.

표 2. 2020년 이후 추가된 속성
Table 2. Attributes added since 2020

External symptom	Internal symptom
Muscle ulcer	Liver degeneration
Eye redness	Liver inflammation
Ocular edema	Spleen liquefaction
Anus hemorrhages	Kidney liquefaction
Anus intestine hernia	Heart hemorrhages

수집한 속성 자료와 넙치의 복수 연관성을 분석하려면 전처리, 학습, 평가, 예측의 단계를 거쳐야 한다. 기계학습 시스템에서 전처리는 핵심 단계 중의 하나로 의미 있는 속성 추출이 필요하다[14]. 질병이나 외부증상, 내부증상, 기타 넙치의 크기나 수온 등의 항목과 복수의 관계를 분석하기 위해 데이터 전처리 과정을 진행한다.

본 연구에서는 파이썬을 사용하여 랜덤 포레스트를 이용한 복수 증상 분석 시스템을 구현하였으므로 파이썬 데이터 프레임(data frame)을 구성하여 단계별로 처리하는 과정을 설명한다. 먼저 수집한 데이터의 종류와 저장 형태, 빈도수를 고려하여 분석에 사용할 항목을 선택하였다. 선택의 기준을 정한 후 향후 다양한 방법으로 분석할 수 있도록 세분화를 진행하였다.

가. 제거할 속성 항목의 선택과 처리

실험에 사용할 질병 진단 데이터는 복수 증상 분석과 연관성이 낮은 텍스트 위주의 항목, 2020년부터 추가되어 2015-2019까지 조사하지 않은 항목, 그리고 수온/크기/무게와 같이 스케일이 달라 정확도에 영향을 주는 항목 등이 있다.

따라서 기계학습에 적합하도록 제거할 속성 항목을 위의 3가지로 구분하고 별도로 테스트를 진행하여 그 결과를 비교하였다. 특히 두 번째 경우로 조사하지 않거나 값이 누락된 속성의 경우에는 항목 자체를 입력 자료로 사용하지 않는 방법과 미측정 기간에 대해 NA로 대치한 후 처리하는 방법으로 구분하고 결과를 확인하였다.

나. 속성 항목의 통합

일부 속성 항목은 속성 간의 상관관계가 높아 중복된 정보를 가지고 있어 저장 공간을 줄이고 학습 알고리즘 실행 속도도 높일 수 있도록 통합을 진행하였다. 이는 학습 데이터셋에 관련이 부족한 속성이 많으면 모델의 예측 성능을 떨어뜨리는 점을 고려한 것이다. 현미경 검경을 통해 검출된 증상과 질병 증상별 다량, 소량, 중량 항목은 지나치게 세분화되어 있어 모델을 단순화하도록 세 개의 속성을 통합하였다. 가장 관심이 있는 복수 항목은 투명과 불투명을 구분하지 않고 복수 항목을 추가하여 그 값으로 구분하도록 정리하였다. 추가된 항목은 기계학습 시스템의 타깃 레이블(target label)로 사용하는데, 두 가지 경우로 테스트하였다. 이 속성과 관련한 내용은 3.3절에서 자세히 정리한다.

다. 측정값이 누락된 속성 항목의 처리

앞의 표 2에서 보여준 10개의 속성이 2020년부터

추가되어, 2015년부터 2019년까지 수집한 11,273개의 데이터는 해당 속성의 값이 모두 존재하지 않는다. 또한 이 외에도 데이터 수집 과정에서 여러 이유로 값이 누락된 속성 항목들이 있다(표 3).

표 3. 누락된 자료의 속성 이름과 자료 갯수
Table 3. Attribute names and counts of missing data

Attribute Name	Count
Fish species	50
Width	196
Weight	401
Disease Diagnose	14,099

측정 값이 누락된 속성 항목을 처리하기 가장 쉬운 방법은 누락된 값을 가진 행이나 열을 삭제하는 것인데 단순하지만 데이터의 손실이 많은 방법이다.

그러나 앞에서 설명한 삭제 방법을 사용할 경우 데이터의 크기가 많이 감소하므로 누락 값을 치환하는 방법이 있다. 누락된 값을 치환하는 방법이 상황에 따라 많이 달라서 자세한 내용은 4.1에서 설명하겠다.

라. 단계별 전처리 및 결과

앞서 설명한 가~다의 세 가지 전처리를 위해 빅데이터 처리와 분석을 위한 전처리 과정에서 가장 널리 사용되는 파이썬 라이브러리인 판다스(Pandas)를 이용하였다.

1단계로 판다스를 이용해 초기 데이터셋으로 20,234 x 252차원의 데이터 프레임(data frame)을 구성하였으며, 2단계는 텍스트 값을 가진 불필요한 6개의 속성을 제거하여 20,234 x 246차원이 되었다. 3단계는 같은 증상에 대해 다량/중량/소량으로 구분된 속성을 하나로 결합하고, 2개로 구분된 복수 투명과 불투명 속성을 하나로 통합하고 속성 값을 0(복수 없음), 1(투명), 2(불투명)로 수정하였다.

이후 단계는 측정값이 누락된 속성 항목을 처리하기 위해 결측치를 포함한 모든 데이터를 삭제하는 경우(표 4)와 결측치를 치환하는 경우(표 5)에 따라 다른 과정을 거친다.

표 4는 결측치를 포함하는 모든 데이터를 삭제하는 방법으로 4단계에서는 넙치를 포함한 자료의 행만 선택하여 18,040 x 134 차원이 되었다. 마지막 5단계에서는 결측치를 포함한 모든 행을 삭제하여 결과적으로 134개의 속성을 가진 7,362개의 데이터가 되었다.

표 4. 결측치 삭제 방법의 단계별 전처리 과정과 결과 차원
Table 4. Step-by-step preprocessing and result dimension of missing value deletion method

Step	Preprocessing	Dimension
4	- Select only rows where the fish species is Olive Flounder	18,040 x 134
5	- Delete rows with missing values	7,362 x 134

표 5는 결측치를 치환하는 과정과 차원의 변화를 보여준다. 4단계에서는 어종 속성이 빈 경우는 넙치로 가정하여 치환한 다음 넙치 데이터만 선택하였고, 5단계에서는 너무 많은 결측치를 가진 속성을 제거한 다음, 마지막으로 누락된 값을 가진 행을 삭제하여 결과적으로 124개의 속성을 가진 18,090개, 즉 18,090 x 124 차원의 데이터가 되었다.

표 5. 결측치 치환 방법의 단계별 전처리 과정과 결과 차원
Table 5. Step-by-step preprocessing and result dimension of the missing value replacement method

Step	Preprocessing	Dimension
4	- Replace missing fish species with Olive Flounder - Select only rows where the fish species is Olive Flounder	18,090 x 134
5	- Delete columns with many missing values - Delete rows with missing values	18,090 x 124

3.3 양식 넙치 데이터셋의 중요 특징 시각화

연구에 사용하고자 하는 양식 넙치 데이터셋은 전처리를 거쳐도 134개의 속성 정보를 포함하고 있고, 이 중 대부분은 0과 1로 이루어져 있어 데이터의 특징을 한눈에 파악하기 어렵다. 따라서 데이터셋에 포

함된 이상치를 감지하고 분포를 확인하기 위해 seaborn 등의 시각화 도구를 이용하였다.

그림 3은 넙치의 주요 수치데이터 간의 관계를 표시한 것으로, 수온(Temp)의 변화와 크기의 변화를 보면 다양한 크기의 넙치 질병 데이터를 계절별로 수집한 것임을 확인할 수 있다.

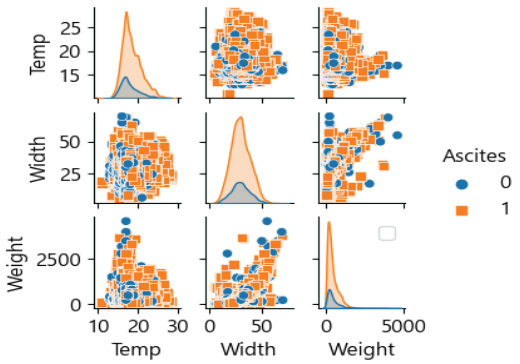

그림 3. 넙치 복수 데이터와 연관 데이터의 pairplot
Fig. 3 Pairplot of Olive Flounder Ascites data and associated features

상관관계 행렬은 속성 간의 선형 상관관계를 표시하므로 속성을 선택하는데 필요한 정보를 제공한다. 그러나 본 연구에서 사용하는 데이터는 속성의 수가 너무 많아 복수(Ascites)와 관련이 깊은 크기(width), 무게(wieght), 비브리오(Vibrio), 그리고 복부 팽만(Abd_dis.)을 선택하여 상관관계를 측정하였다.

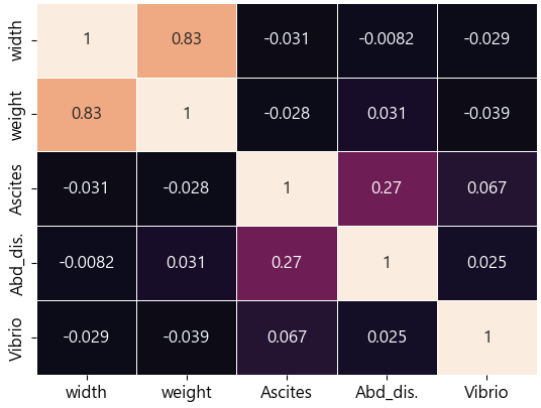

그림 4. 상관관계 행렬 히트맵
Fig. 4 Correlation matrix heatmap

그림 4를 보면 무게와 크기(0.85), 복부 팽만과 복수(0.27)는 상관관계가 비교적 높은 것을 알 수 있다.

또한, 본 연구에서 사용한 18,090개의 데이터셋에서 수치데이터(수온, 크기, 무게)와 타깃 레이블로 사용한 복수 데이터에 대한 간단한 평균(mean), 표준편차(std.), 최솟값(min.)과 최댓값(max.) 그리고 25%, 50%, 75%를 차지하는 통계 정보는 표 6과 같다.

표 6. 사용된 수치 데이터의 간단한 통계 정보
Table 6. Simple statistical information from the numerical data used

	temp.	width	weight	Ascites
mean	18.24	29.74	402.30	0.65
std.	2.40	9.90	390.08	0.48
min.	0.00	0.00	0.00	0.00
25%	16.50	23.00	135.00	0.00
50%	17.60	29.00	280.00	1.00
75%	19.70	36.00	545.00	1.00
max.	28.40	335.00	4000.00	1.00

3.4 복수 증상 관련 데이터 정리

넙치의 복수 여부와 다른 속성들과의 연관성을 찾기 위해 먼저 복수를 학습 모델의 클래스 레이블로 선택한다. 보통 클래스 레이블은 싸이킷런의 분류 추정기에 정수형 배열로 전달하는데 본 연구에서는 'IS_복수_투명', 'IS_복수_불투명' 두 가지 속성치를 더해서 복수 속성 항목으로 추가했다. 추가된 복수 속성이 기존 두 개의 속성 내용을 그대로 반영하므로 분석 과정에서는 두 개의 속성은 삭제하였다. 표 7은 분석에 사용한 복수 자료 개수를 보여준다.

표 7. 복수 증상의 속성값에 따른 자료 건수
Table 7. Number of data according to attribute values of ascites symptoms

Value	Ascites symptom	Count
0	none	7,816
1	transparent	2,421
2	opaque	9,997
Total		20,234

본 연구에서는 타깃 레이블인 복수를 표 8과 같이 A, B, C 세 가지 경우로 나누어 분석한다. Case A는 복수 증상이 없으면 0, 투명이나 불투명이면 1을 가지며, Case B는 복수 증상이 없으면 0, 투명이면 1, 불투명이면 2를 갖는다. 마지막 Case C는 복수 증상 없음을 제외하고 투명이면 0, 불투명이면 1을 갖는다.

표 8. 각각의 Case에 사용된 타깃 레이블
Table 8. Target label used each case

Case	none	transparent	opaque
A	0	1	
B	0	1	2
C	-	0	1

Ⅳ. 실험 및 결과 분석

4.1 누락 자료 처리

Ⅲ 장에서 살펴본 바와 같이 원본 데이터는 다수의 칼럼에 누락 자료를 포함하고 있어 해당 데이터 전체를 삭제하면 데이터의 크기는 5,098,968(20,234 x 252) 차원에서 233,820(7,362 x 135) 차원으로 약 95%가 줄어들게 된다. 이 중 실제 누락 자료 항은 129,320건이므로 하나의 항이라도 누락 자료를 가지면 해당 행을 전부 삭제하는 방법은 분석을 위해 수집한 자료 대부분을 사용할 수 없게 되는 문제가 발생한다. 이와 같은 과도한 샘플 삭제를 막기 위한 누락 자료 치환에 대해 설명한다.

가. 누락 값이 적은 열의 값 치환

크기, 무게 속성은 상대적으로 누락 건수가 적으므로 데이터의 분포를 고려해서 평균값으로 치환한다. 어종 속성은 넙치에 대한 자료를 수집하는 중 기타 어종 자료도 추가한 상황을 고려하면 종을 표시하지 않고 샘플 데이터를 작성한 경우는 넙치에 대한 자료라 판단할 수 있다. 따라서 누락 자료 50건의 어종 행은 넙치로 치환한다. 이 과정을 통해 원본 데이터의 약 20%를 학습에 사용할 수 있게 되었다.

나. 누락 값 개수가 기준점 이상인 열 삭제

2020년 추가한 10개 속성은 누락 행이 11,273개로 50% 이상의 샘플 자료가 측정되지 않은 경우이다. 이 속성을 남겨두고 누락 자료 행을 삭제한다면 2020년 이전의 자료는 분석에서 제외한 것과 마찬가지인 상황이 발생한다. 따라서 이 10개의 속성에 해당하는 열을 삭제하면 원본 자료의 약 44%를 학습 대상 자료로 이용할 수 있다. 본 연구에서는 결측치를 치환한 후 테스트를 진행하였으며, 치환하지 않은 경우와 그 결과를 비교하였다.

4.2 데이터셋 분할과 스케일 조정

모델 학습과 테스트 및 성능 평가를 위해서 데이터셋을 훈련 세트와 테스트 세트로 나눈다. 본 연구에서는 분석할 데이터셋의 속성 레이블은 X, 타깃 클래스 레이블은 y에 각각 저장한다. 사이킷런(sklearn)의 train_test_split 함수를 사용해서 80%는 훈련 데이터, 20%는 테스트 데이터로 랜덤하게 나누었다. 데이터셋은 훈련과 테스트 셋으로 분할하기 전에 무작위로 섞고, 계층화 기능을 사용하여 훈련 세트와 테스트 세트의 클래스 레이블 개수가 입력 데이터셋과 비율이 같도록 구성하였다. 성능 향상을 위해 사이킷런의 StandardScaler를 사용하여 속성 스케일을 조정하였다. 모델 학습과 테스트를 위한 타깃 레이블은 y_train, y_test로 표시한다. 복수의 유/무, 투명/불투명 등의 구분에 따른 A, B, C에서 데이터셋을 분할한 수는 표 9와 같다.

표 9. 훈련과 테스트 데이터셋의 분할
Table 9. Split of training and test data set

Case	Target label	Ascites	y_train	y_test
A	0 (none)	6,389	5,111	1,278
	1 (exist)	11,701	9,361	2,340
	subtotal	18,090	14,472	3,618
B	0 (none)	6,389	5,111	1,278
	1 (transparent)	2,274	1,819	455
	2 (opaque)	9,427	7,542	1,885
	subtotal	18,090	14,472	3,618
C	0 (transparent)	2,421	1,937	484
	1 (opaque)	9,997	7,997	2,000
	subtotal	12,418	9,934	2,484

4.3 분류 알고리즘

이 절에서는 분류모델과 최적화 알고리즘을 선택하고 모델을 평가하는 단계를 설명한다. 복수의 있음과 없음을 구별하는 본 연구의 모델에는 분류 알고리즘이 적합하여 분류 알고리즘 중 랜덤 포레스트를 선택하고 모델 학습과 검증을 진행하였다. 유용한 속성을 선택하는 랜덤 포레스트는 결정트리에서 계산한 불순도를 이용하여 속성의 중요도를 측정한다. 본 연구의 입력 데이터셋에는 252개의 속성이 있는데, 먼저 관계없는 속성이나 잡음을 제거하고 모델을 단순하게 구성하는 방법을 찾기 위해 복수와 연관성이 높은 중요한 속성을 찾았다.

가. 랜덤 포레스트 학습

양식 넙치의 복수를 A, B, C 경우 각각에 대해 수조 수온(Temperature), 넙치 크기(Width), 넙치 무게(Weight), 외부증상(ES: External clinical signs), 내부증상(IS: Internal clinical signs), 현미경 검경(Micro related symptoms)과 병원체(Pathogens) 정보를 독립변수로 하여 분석을 수행하였다. 또한 현미경 검경(Micro related symptoms)을 통한 질병 진단 데이터는 다량, 중량, 소량을 하나의 속성으로 통합하여 분석하였다. 2020년도 이전에는 조사하지 않아 누락된 부분이 많은 속성에 대해서는 결측치를 치환하였다.

넙치의 복수 증상과 기타 속성들과의 연관성을 찾기 위해 case A(복수 없음/있음)와 case B(복수 없음/투명/불투명), case C(투명/불투명)의 세 가지로 나누고, 각각을 속성 타깃 레이블로 지정하여 테스트를 진행한 후 모델을 평가하였다.

나. 유용한 속성 선택과 차원 축소

모델의 복잡도를 줄이는 속성 선택은 원본이 가지고 있는 속성들에서 일부를 선택하여 과대적합을 피하는 방법으로 주어진 문제를 해결하기 위해 가장 관련성이 높은 속성의 부분 집합을 고르는 것이다.

랜덤 포레스트에서는 모델을 훈련한 후 각 속성의 중요도를 바탕으로 사이킷런의 SelectFromModel 클래스를 사용하여 중요 속성을 선택하였다. 중요 속성 선택은 중요도를 선택하는 기준점(threshold) 값을 매개변수로 모델에 전달하고 기준점 이상인 중요도를 가지는 속성을 골라 입력 속성의 차원을 축소하는 것이다.

그림 5는 Case A의 데이터를 이용하여 학습된 랜덤 포레스트 모델에서 기준점이 0.01일 때 14,472개의 샘플에서 선택된 36개의 속성의 중요도와 누적치를 그래프로 표현한 것이다. 속성의 이름을 모두 표현하기에는 한계가 있어 x축에서 속성의 이름 대신 중요도가 높은 속성의 순위를 표시하였다. 선택된 36개 속성들의 중요도 합은 약 80%이다. 기준점이 0.015이면 25개, 0.02이면 13개의 속성이 선택되었다. 상위 10개의 속성 이름과 중요도는 표 10의 Case A에서 확인할 수 있다.

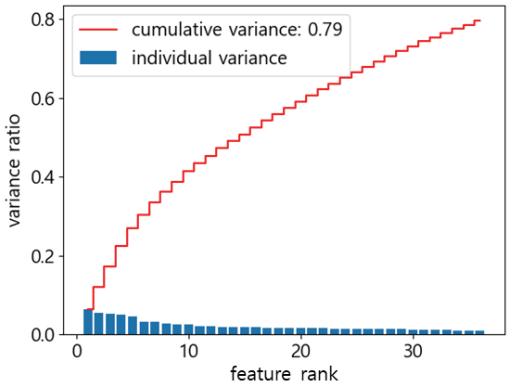

그림 5. Case A의 속성 중요도
Fig. 5 Important features of Case A

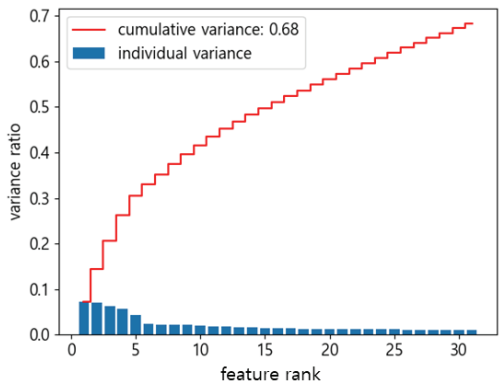

그림 6. Case B의 속성 중요도
Fig. 6 Important features of Case B

그림 6은 Case B의 데이터를 이용하여 학습된 랜덤 포레스트 모델에서 기준점이 0.01일 때 14,472개

표 10. 복수 증상과 관련한 3가지 Case 별로 상위 10개의 주요 속성 비교
Table 10. Comparison of the top 10 key attributes for each of the three cases related to ascites symptoms

	Case A		Case B		Case C	
Rank	Attribute name	Importance	Attribute name	Importance	Attribute name	Importance
1	Temperature	0.064	Abdomen distension	0.072	Abdomen distension	0.183
2	Weight	0.055	Temperature	0.071	Temperature	0.069
3	Gill pale	0.053	Weight	0.062	Weight	0.068
4	Width	0.051	Width	0.057	Width	0.064
5	Abdomen distension	0.045	Gill pale	0.043	Anus intestine hernia	0.038
6	Argulus japo nicus	0.033	MB_Scuticociliatida	0.024	Heart enlargement	0.023
7	MB_Scuticociliatida	0.033	Liver enlargement	0.023	Intraperitoneal muscle petechial hemorrhages	0.019
8	Heart enlargement	0.027	Argulus japo nicus	0.022	Skin petechial hemorrhage	0.017
9	MI_Vibrio spp.	0.026	Heart enlargement	0.022	Spleen nodule	0.017
10	Liver enlargement	0.025	MI_Vibrio spp.	0.019	Intraperitoneal muscle hemorrhages	0.016

의 샘플에서 선택된 31개의 속성의 중요도와 누적치를 그래프로 표현한 것이다. 선택된 31개 속성들의 중요도 합은 약 68%이다. 이중 상위 10개 속성의 이름과 중요도는 표 10의 Case B에서 확인할 수 있다.

그림 7은 Case C의 데이터를 이용하여 학습된 랜덤 포레스트 모델에서 기준점이 0.01일 때 9,934개의 샘플에서 선택된 27개의 속성의 중요도와 누적치를 그래프로 표현한 것이다. 선택된 27개 속성들의 중요도 합은 약 73%이다.

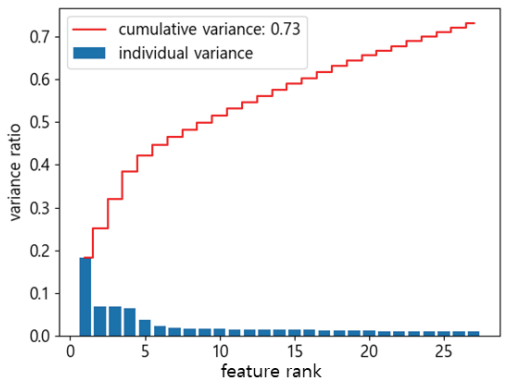

그림 7. Case C의 속성 중요도
Fig. 7 Important features of Case C

이중 상위 10개 속성의 이름은 “ES_복부_팽만, 수온, 무게, 크기, ES_항문_탈장, IS_심장_비대, ES_복강내근육_점상출혈, ES_체표_점상출혈, IS_비장_결절, ES_복강내근육_출혈”이며, 각각의 중요도는 표 10의 Case C에서 확인할 수 있다.

4.4. 모델 학습 결과 비교

표 10은 복수 증상 관련한 3가지 Case 각각에 대해 중요 속성으로 선택된 상위 10개의 속성을 비교한 것이다. Case A(복수 없음/있음)와 B(복수 없음/투명/불투명) 모델의 경우 중요도에 따른 순서의 차이가 있지만 상위 10개의 속성은 동일하다. 이는 Case A와 Case B는 동일한 데이터 셋으로 이진 분류와 다중 분류 모델을 구성했기 때문이다. 반면에 Case C(복수 투명/복수 불투명)의 경우는 복수 없음 데이터를 배제하고 이진 분류 모델을 구성했기 때문에 학습에 사용한 데이터 셋의 개수에 차이가 있어서 약간 다른 속성이 나온 것으로 판단된다.

4.2절의 표 8에 제시한 바와 같이 Case A(복수 없음/있음), B(복수 없음/투명/불투명), C(복수 투명/복수 불투명) 각각에 대해 훈련 데이터를 나누고, 랜덤 포레스트 분류 모델을 학습시켰다. 훈련된 모델의 정확도를 측정하기 위해 검증 데이터를 predict 메소드에 전달하여 모델이 분류한 예측값을 저장하고, 예측값을 실제 데이터와 비교하여 분류모델의 예측 정확도를 평가하는 지표인 오차 행렬(Confusion Matrix)을 계산하였다[15].

표 11은 원래의 전체 속성을 그대로 사용한 모델과 임계값 0.01을 기준으로 속성의 차원 축소를 적용

한 모델로 나누어 이진 분류(Case A와 C)와 다중 분류모델(Case B)의 학습 정확도와 오차 행렬을 비교한 결과이다. 차원을 축소해도 정확도가 크게 떨어지지 않음을 보여준다.

표 11. 랜덤 포레스트 학습 결과 비교
Table 11. Comparison of Random forest learning results

Case	Comparision criteria	Original all attributes	Down sampled attributes
A	dimension	(18,090 x 124)	(14,472 x 36)
	accuracy	0.90	0.86
	confusion matrix	114 137 212 2,128	1,071 207 284 2,056
B	dimension	(18,090 x 124)	(18,090 x 31)
	accuracy	0.83	0.78
	confusion matrix	1,154 5 119 25 263 167 218 90 1,577	1,087 6 185 22 259 174 270 124 1,491
C	dimension	(12,418 x 124)	(9,934 x 27)
	accuracy	0.89	0.89
	confusion matrix	278 206 77 1,923	294 190 117 1,883

4.5. 분류 모델의 검증

본 연구에서 구현한 분류 문제는 Case A(복수 없음/있음), B(복수 없음/투명/불투명), C(복수 투명/불투명)로 Case A와 C는 이진 분류 모델이고, Case B는 다중 분류 모델이다. 따라서 이진 분류 모델과 다중 분류 모델을 평가하는 기준은 차이가 있다. 여기에서는 이진 분류모델인 Case A와 C는 ROC 곡선으로 검증하였으며(그림 8, 그림 9), 다중 분류모델인 Case B는 표 12와 같이 metrics 모듈의 classification_report 함수를 사용하여 정밀도(Precision), 재현율(Recall), F1-score 지표로 모델의 예측 능력을 확인하였다.

ROC(Receiver Operating Characteristic) 그래프는 오차 행렬의 TPR(True Positive Rate)과 FPR(False Positive Rate) 점수를 기반으로 분류모델의 성능을 확인할 수 있는 도구이다. 그래프에서 ROC 곡선의 아래 면적 ROC AUC(ROC Area Under the Curve)를 계산하여 케이스별 성능을 확인할 수 있으며 이진 분류 모델의 성능을 평가하는데 주로 이용된다.

그림 8과 그림 9에서 파란색 점선으로 표시된 완벽한 분류기일 경우 TPR이 1이고 FPR이 0이며, 빨간색의 점선으로 표시된 dummy 분류기의 경우 랜덤 추측인 경우를 표시한다. 본 연구의 데이터를 이용한 곡선은 전체 속성을 모두 사용한 오렌지색의 랜덤 포레스트 모델(Random Forest)과 중요 속성을 선택하고 훈련한 초록색으로 그려진 랜덤 포레스트 모델(SFM Random Forest)이다. 두 가지 경우 모두 ROC AUC는 1.0과 0.5 사이에 위치하고 있음을 확인하였다.

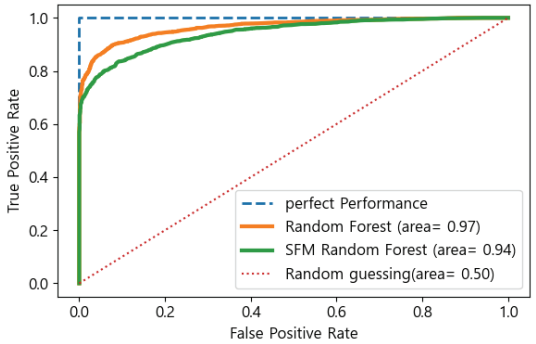

그림 8. Case A의 ROC 곡선
Fig. 8 ROC curve of Case A

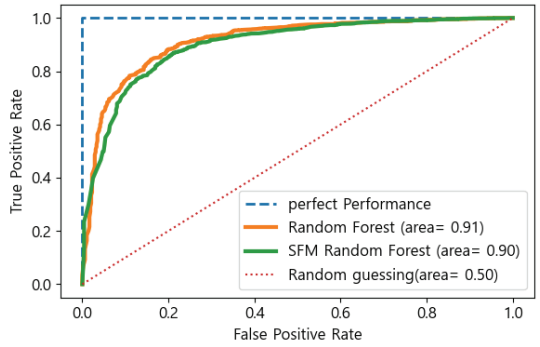

그림 9. Case C의 ROC 곡선
Fig. 9 ROC curve of Case C

표 12는 Case B(복수 없음/투명/불투명)의 다중 분류 모델의 성능을 정밀도(Precision)와 재현율(Recall), F1-score 지표를 측정해서 보여준다. 각각

의 지표는 다음 식 (1) ~ (3)으로 계산된다. 식에서 TP, FP, FN은 각각 True Positive, False Positive, False Negative를 나타낸다.

$$\mathrm{Precision} = \frac{TP}{TP + FP} \quad \cdots (1)$$

$$Recall = \frac{TP}{TP + FN} \quad \cdots (2)$$

$$F1 - score = \frac{2*(\mathrm{Precision} * Recall)}{\mathrm{Precision} + Recall} \quad \cdots (3)$$

표 12(a)는 모든 속성을 사용했을 때의 성능을 보여주고, (b)는 중요도가 0.01 이상인 속성만 사용했을 때의 성능이다. 랜덤 포레스트 모델을 이용해 속성의 차원을 축소한 경우에도 전체 차원을 그대로 사용한 경우에 비교하여 성능이 크게 떨어지지 않음을 확인할 수 있다.

표 12. Case B의 분류 성능 비교
Table 12. Comparison of classification performance of Case B

Class	Precision	Recall	F1-score
0	0.83	0.90	0.86
1	0.73	0.58	0.65
2	0.85	0.84	0.84

(a) 모든 속성을 사용한 경우
(a) When all features are used

Class	Precision	Recall	F1-score
0	0.79	0.85	0.82
1	0.67	0.57	0.61
2	0.81	0.79	0.80

(b) 선택된 특성을 사용한 경우
(b) When selected features are used

4.6 실험 결과 고찰

복수 증상 데이터 분석을 위한 전처리 과정으로써 2015년부터 수집된 7년간의 질병 진단 데이터에서 관련이 적은 텍스트 항목의 삭제, 새로 추가되어 기존에 수집한 데이터에 전혀 존재하지 않거나 기타의 이유로 대량 누락된 속성 항목의 삭제, 동일한 증상에 대해 다량/중량/소량 등으로 세분한 속성 항목을 통합하고, 측정값이 누락된 항목에 대해 삭제하거나, 치환 등의 작업을 수행하였다.

다음으로 복수 증상과 관련하여 Case A(복수 없음/있음), B(복수 없음/투명/불투명), C(복수 투명/불투명) 3가지 경우로 구분하고 랜덤 포레스트 모델을 학습하여 복수 증상과 관련이 있는 일정 임계값 이상의 주요 속성을 추출하였다.

추출된 주요 속성을 이용해 3가지 Case의 분류 성능을 측정하였으며, 전체 속성을 모두 사용할 때의 분류 성능과 비교하였다. 비교 항목으로는 이진 분류 문제인 Case A와 Case C는 ROC AUC, 3항 분류 문제인 Case B는 정밀도, 재현율, F1-Score을 사용하였다.

그림 8과 그림 9, 표 12의 결과에서 알 수 있듯이 랜덤 포레스트 모델을 통해 복수 증상과 관련 있는 주요 속성들을 추출할 수 있음을 확인하였다. 이는 향후 질병 진단 전문가나 양식 어민이 넙치를 직접 해부하지 않더라도 복수 증상과 관련 있는 일부의 증상만으로 복수 관련 질병을 예측하는 데 활용할 수 있다는 장점이 있다.

그러나, 분류 성능이 다른 분류 문제보다 떨어지는데 이는 랜덤 포레스트 모델에 사용한 질병 진단 데이터가 너무 많은 증상 항목에 대해 단순히 증상의 유/무로만 기록이 되어 있어 설명력이 부족하기 때문인 것으로 판단된다.

Ⅴ. 결론 및 향후 연구

본 논문에서는 양식장의 생산력에 영향을 미치는 복수 증상에 대한 집중적인 분석을 위해, 복수 증상 속성을 Case A(복수 없음/있음), B(복수 없음/투명/불투명), C(복수 투명/불투명)으로 구분하고 파이썬의 라이브러를 활용해 랜덤 포레스트 모델을 구축하였다.

이 모델을 통해 복수 증상과 관련이 있는 일정 임계값 이상의 주요 속성을 추출하였으며, 추출된 주요 속성 데이터만 사용하여 3가지 Case의 분류 성능을 측정한 결과, 전체 속성 데이터를 사용한 경우와 가까운 성능을 보였다. 즉, 랜덤 포레스트 모델을 통해 복수 증상과 관련 있는 주요 속성들을 추출할 수 있음을 확인하였다. 이는 향후 질병 진단 전문가나 양식 어민이 넙치를 직접 해부하지 않더라도 복수 증

상과 관련 있는 일부의 증상만으로 복수 관련 질병을 예측하는 데 활용할 수 있다는 장점이 있다.

그러나, 분류 성능이 다른 분류 문제보다 떨어지는데 이는 랜덤 포레스트 모델에 사용한 질병 진단 데이터가 너무 많은 증상 항목에 대해 단순히 증상의 유/무로만 기록이 되어 있어 설명력이 부족하기 때문으로 보인다. 이는 추후, 양식 넙치의 질병 진단 데이터를 기록할 때, 증상의 정도(상, 중, 하) 또는 세균이나 박테리아의 양 등을 수치로 표현한다면 성능이 개선될 것으로 보인다.

감사의 글

본 논문은 2023년 해양수산부 재원으로 해양수산과학기술진흥원의 지원을 받아 수행된 연구임 (스마트 수산양식 연구센터).

References

[1] H. Kim, S. Jung, S. Kim, J. Park, H. Ceong, and S. Han, “Data Mining for Scuticociliatosis Outbreak Patterns in Cultured Olive Flounder Paralichthys olivaceus in Jeju, Korea,” *Korean J. Fisheries and Aquattic Sciences*, vol. 53, no. 5, 2020, pp. 740-751.

[2] K. Kim, S. Han, T. Kim, S. Jung, S. Kim, H. Ceong, and J. Park, “Pattern Analysis of Clinical Signs in Cultured Olive Flounder, *Paralichthys Olivaceus,* with Edwardsielosis using the Decision Tree Technique,” *J. of the Korea Institute of Electronic Communication Sciences*, vol. 16, no. 4, 2021, pp. 661-674.

[3] A. Cutler, D. Cutler, and J. Stevens, “Random forests,” *Machine Learning*, vol. 45, no. 1, 2011, pp. 157-176.

[4] S. Zhao, S. Zhang, J. Liu, H. Wang, and R. Zhao, “Application of machine learning in intelligent fish aquaculture: a review,” *Aquaculture*, vol. 540, 2021, pp. 724-736.

[5] V. Lyubchenko, R. Matarneh, O. Kobylin, and V. Lyashenjo, “Digital image processing techniques for detection and diagnosis of fish diseases,” *Int. J. of Advanced Research in Computer Science and Software Engineering*, vol. 6, no. 7, 2016, pp. 79-83.

[6] S. Malik, T. Kumar, and A. K. Sahoo, “A novel approach to fish disease diagnostic system based on machine learning,” *Advances in Image and Video Processing,* vol. 5, no. 1, 2017, pp. 49-57.

[7] M. Ahmed, T. Aurpa, and M. Azad, “Fish disease detection using image based machine learning technique in aquaculture,“ *J. King Saud Univ. Comp. Informat. Sci.*, vol. 34, 2022, pp. 5170-5182.

[8] H. Son, H. Lim, and H. Choi, “A Study on Disease Prediction of Paralichthys Olivaceus using Deep Learning Technique,“ *Smart Media J.*, vol. 11, no. 4, 2022, pp. 62-68.

[9] J. Li, Z. Lian, Z. Wu, L. Zeng, L. Mu, Y. Yuan, H. Bai, Z. Guo, K. Mai, X. Tu, and J. Ye, “Artificial intelligence‐based method for the rapid detection of fish parasites (Ichthyophthirius multifiliis, Gyrodactylus kobayashii, and Argulus japonicus),” *Aquaculture*, vol. 563, Part 1, 2023, Article ID 738790, pp. 1-9.

[10] M. Khan, A. Qayoom, M. Nizami, M. Siddiqui, S. Wasi, and S. Raazi, "Automated Prediction of Good Dictionary EXamples (GDEX): A Comprehensive Experiment with Distant Supervision, Machine Learning, and Word Embedding-Based Deep Learning Techniques," *J. Complexity*, vol. 2021, Article ID 2553199, 2021, pp. 1-18.

[11] M. Islam, M. Kashe, and J. Uddin, “Fish survival prediction in an aquatic environment using random forest model,” *IAES Int. J. of Artificial Intelligence (IJ-AI)*, nol. 10, no. 3, 2021, pp. 614-622.

[12] J. Luan, C. Zhang, B. Xu, Y. Xue, and Y. Ren, “The predictive performances of random forest models with limited sample size and different species traits,” *Fisheries Research,* vol. 227, Article ID 105534, 2020, pp. 1-10.

[13] P. Swetha, A. H. K. P. Rasheed and V. P.

Harigovindan, "Random Forest Regression based Water Quality Prediction for Smart Aquaculture," *4th Int. Conf. on Computing and Communication Systems (I3CS)*, Shillong, India, 2023, pp. 1-5.

[14] J. Kang, J. Park, S. Han, and K. Kim, "Development of Machine Learning based Flood Depth and Location Prediction Model," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 18, no. 1, 2023, pp. 91-98.

[15] M. Park, H. Yoon, N. Kim, B. Kim, and H. Yoon, "A Comparative Study on Machine Learning Models for Red Tide Detection," *J. of the Korea Institute of Electronic Communication Sciences*, vol. 16, no. 6, 2021, pp. 1363-1371.

저자 소개

김경임(Kyeong-Im Kim)

1993년 전남대학교 컴퓨터공학과 졸업(공학사)
2013년 전남대학교 디지털컨버전스협동과정 졸업(이학석사)
2019년~현재 스마트수산양식연구센터 연구원
※ 관심분야 : 빅데이터 분석, ICT융합

김성현(Sung-Hyun Kim)

2007년 부경대학교 수산생명의학과 졸업(이학사)
2009년 노르웨이수의과학대학 대학원 Aquatic Medicine학과 졸업(이학석사)
2015년 노르웨이생명과학대학교 대학원 Aquatic Medicine 졸업(이학박사)
2015년~현재 ㈜피쉬케어 대표이사 겸 연구소장
2018년~현재 제주특별자치도 보건환경위원
※ 관심분야 : 수산양식, 수산질병, 스마트양식

한순희(Soonhee Han)

1983년 경북대학교 전자공학과 졸업(공학사)
1985년 광운대학교 전자계산학과 졸업(이학석사)
1993년 광운대학교 전자계산학과 졸업(이학박사)
1998년~현재 전남대학교 문화콘텐츠학부 교수
※ 관심분야 : 이동통신, 임베디드시스템, ICT융합

정희택(Hee-Taek Ceong)

1992년 2월 전남대학교 전산통계학과 학사
1995년 2월 전남대학교 전산통계학과 석사
1999년 8월 전남대학교 전산통계학과 박사
1999년~현재 전남대학교 문화콘텐츠학부 교수
※관심분야 : 데이터마이닝, 기계학습, 분산처리시스템, 빅데이터 분석

박정선(Jeong-Seon Park)

1992년 충북대학교 컴퓨터과학과 졸업(이학사)
1994년 충북대학교 전산학과 졸업(이학석사)
2005년 고려대학교 컴퓨터학과 졸업(이학박사)
1994년~1999년 현대정보기술 선임연구원
2005년~현재 전남대학교 문화콘텐츠학부 교수
※ 관심분야 : 컴퓨터비전, 영상처리, ICT융합