이중 스트림 네트워크 기반 난임 시술 데이터셋을 활용한 난자 채취 결과 예측 모델

이재원 [1]*, 이상범 [1], 김범휘 [1], 박효진 [2], 양슬기 [2], 구덕본 [2], 김광용 [1]
*[1] 한국전자통신연구원, [2] 대구대학교

*jwl0127@etri.re.kr,

A prediction model for oocyte retrieval results using a dual-stream network-based infertility treatment dataset

Lee Jaewon [1]*, Lee Sangbeom[1], Kim Bumhwi[1], Park Hyo-Jin[2], Yang Seul-Gi[2], Koo Deog-bon[2], Kim Kwang-Yong[1]
*[1]Electronics and Telecommunications Research Institute (ETRI), [2]Daegu Univ.

요 약

본 논문은 이중 스트림 네트워크를 활용한 난임 시술을 위한 채취 난자 수를 예측하는 모델을 제안한다. 난임 시술에 있어 채취 난자 수는 시술의 종류, 처리 방법, 난자 동결 등 난임 시술과 관련된 선택지를 결정하는데 결정적인 지표이다. 이를 예측하여 정보를 제공해주는 방법이 필요하지만 일반적인 관련 데이터 셋은 범주형 데이터와 수치형 데이터가 혼합 되어있어 서로 다른 데이터 형태가 혼재되어 효과적인 학습이 어렵다. 따라서 본 논문에서는 이중 스트림 네트워크 구조를 활용하여 범주형과 수치형 데이터로부터 각 데이터로부터 정보를 추출한 후 통합하여 채취 난자 수를 예측하는 모델을 제안한다. 제안한 모델을 활용하여 11 종의 혼합 데이터 필드 기반 예측한 결과 0.7235 의 상관관계를 보이는 것을 확인하였고, 난자 수 8 개를 기준으로 이진분류한 결과 79.2 %의 Accuracy 를 보였다. 또한 이를 전통적인 방식과 비교한 결과 제안한 방법이 가장 개선된 상관관계를 보이는 것을 확인하였다. 이런 결과를 바탕으로, 제안한 예측 모델을 통해 채취 난자 수를 더욱 정확하게 예측할 수 있고 이를 활용하여 추후 난임 시술을 준비하는 사람들에게 사전적인 정보를 제공해 줄 수 있는 수단으로 활용 가능할 것으로 예측한다.

Ⅰ. 서 론

난임 시술에서 채취 난자 수는 시술의 방향과 결과를 결정짓는 핵심 지표로 활용된다 [1]. 예를 들어, 채취된 난자 수에 따라 배아 배양 전략, 동결 여부, 추가 시술 방법 등이 달라진다. 따라서 환자 맞춤형 시술 계획을 수립하기 위해 사전에 난자 수를 예측할 수 있는 모델의 필요성이 꾸준히 제기되고 있다 [2]. 그러나 실제 임상 데이터는 다양한 범주형 변수(호르몬 사용 여부, 시술 방법 등)와 수치형 변수(연령, 체질량지수, 항뮬러관 수치 등)가 혼재되어 있으며, 이로 인해 예측 모델의 학습 성능이 제한되는 문제가 존재한다. 기존 연구에서는 단일 형태의 입력 데이터에 최적화된 모델이 주로 사용되어, 복합적 데이터의 효과적 통합이라는 과제가 남아 있다 [3].

Ⅱ. 본론

본 연구에서는 69,255 건의 실제 난임 시술 데이터를 기반으로 총 11 종의 입력 변수를 활용하였다. 데이터는 환자의 연령, 체질량지수(BMI), 시술 횟수와 같은 수치형 변수와, 배란유도 방법, 호르몬 종류, 보조제 사용 여부와 같은 범주형 변수를 포함한다. 원본 데이터는 일부 결측값과 이상치를 포함하여 데이터 정제 및 표준화, 그리고 범주형 변수에 대한 원-핫 인코딩을 수행하였다. 이러한 전처리 과정을 통해 모델 학습에 적합한 입력 형태로 데이터를 정제하였다. 정제 결과, 수치형 데이터 5 종과 범주형 데이터 42 종으로 구성되었고 이 데이터를 활용하여 제안한 모델을 학습하였다. Training set 과 Test set 은 4 대 1 의 비율로 random split 하여 실험을 진행하였고, 모델의 최적화를 위한 validation set 은 training set 의 10%로 배정하여 활용하였다.

본 연구는 대구대학교 기관윤리위원회 (IRB No. 1040621-202505-HR-032)의 승인을 받았으며, 사전 동의 요건은 면제되었으며 개인정보는 익명화 후 제공되었습니다.

제안한 모델은 이중 스트림 네트워크 (Dual-stream Network) 구조를 기반으로 한다. 첫 번째 스트림은 범주형 변수를 입력받아 임베딩과 다층 퍼셉트론 층을 통해 특성을 추출하며, 두 번째 스트림은 수치형 변수를 입력받아 연속형 특성을 학습한다. 이후 두 스트림의 출력을 통합하여 난자 채취 수를 예측하는 최종 회귀 및 분류 모듈로 연결하였다. 이러한 구조를 통해 서로 다른

데이터 특성이 주어진 데이터셋으로부터 적합한 표현을 학습하고, 통합된 정보로부터 보다 정확한 예측이 가능하다.

예측 성능 평가는 두 가지 방식으로 진행되었다. 첫째, 채취 난자 수를 연속형 변수로 예측하여 상관계수(Correlation coefficient)를 산출하였고, 둘째, 기준 난자 수를 8 개로 설정하여 이진 분류를 수행하였다. 그 결과, 제안한 모델은 난자 수 예측에서 0.7235 의 상관계수를 보였으며, 이진 분류에서는 79.5%의 F1-score 를 달성하였다.

제안한 모델의 성능을 검증하기 위해 두 가지 비교 모델을 설정하였다. 첫 번째는 범주형과 수치형 변수를 구분하지 않고 하나의 입력으로 결합하여 학습하는 단일 입력 DNN 모델이며, 두 번째는 범주형과 수치형 데이터를 함께 처리할 수 있는 전통적 기계학습 방법인 Decision Tree 모델이다. 그 결과는 아래 표 1 에서 볼 수 있다. 단일 입력 DNN 모델은 상관계수 0.7011, Accuracy 77.8%를 보였으며, Decision Tree 모델은 상관계수 0.6913, Accuracy 76.5%로 나타났다. 이를 통해 범주형과 수치형 변수를 독립적으로 처리한 후 통합하는 전략이 예측 정확도를 높이는 데 효과적임을 확인하였다.

표 1. 제안 모델과 기존 모델간 성능 비교

Model	Coefficient	Accuracy
Decision Tree	0.6913	76.5
Single DNN	0.7011	77.8
Dual stream DNN	0.7033	79.2

또한 모델의 예측에 대한 해석을 위해 Shapley value [4]에 기반한 데이터 feature importance 와 실제 예측과 데이터 값의 변화를 살펴보기 위해 아래 그림 1 과 같이 확인해보았다. 그 결과 AMH 수치가 가장 높은 중요도를 보였고, 그 이후 난임 시술 시도 횟수, 호르몬 사용, 나이 순의 결과를 보였다. 또한 Shap value 의 변화를 값의 크기에 따라 표현되어 있는데 해당 데이터가 높을수록(붉은색) 예측 값이 커지는 (기준선 오른쪽 배치) 경향이 보이는 것을 확인할 수 있다.

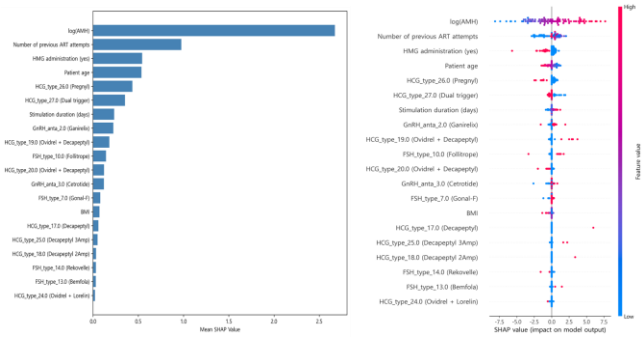

그림 1. SHAP value 기반 feature importance 분석 결과

Ⅲ. 결론

본 연구는 난임 시술 데이터의 복합적 특성을 고려하여 이중 스트림 네트워크 기반 예측 모델을 제안하였으며, 이를 통해 난자 채취 수를 기존 방식보다 높은 정확도로 예측할 수 있음을 확인하였다. 특히, 범주형과 수치형 데이터를 효과적으로 통합하는 전략은 향후 난임 데이터 기반 AI 연구의 확장 가능성을 보여준다. 제안된 모델은 임상 현장에서 환자 맞춤형 시술 전략을 수립하는 데 기여할 수 있으며, 더 나아가 환자에게 사전적 정보를 제공함으로써 시술 과정에서의 불확실성을 줄이는 데 도움이 될 것으로 기대된다.

.

ACKNOWLEDGMENT

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by the Korean government [25ZD1140, Regional Industry ICT Convergence Technology Advancement and Support Project in Daegu-GyeongBuk (Medical); 25ZD1170, Regional Industry ICT Convergence Technology Advancement and Support Project in Daegu-GyeongBuk (Divison)]

참 고 문 헌

[1] Esteva A, Robicquet A, Ramsundar B, et al. A guide to deep learning in healthcare. Nat Med. 2019; 25(1):24-29.

[2] Liang X, Huang J, Li C, Sun W, Liu J. Machine learning-based prediction of the number of oocytes retrieved in IVF treatment. Reprod Biomed Online. 2021; 42(3):580-588.

[3] Sakkas D, Gardner DK. Artificial intelligence in the in vitro fertilization (IVF) laboratory: a review of current status and future prospects. Fertil Steril. 2023;120(2):233-244.

[4] Lundberg SM, Lee SI. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems (NeurIPS). 2017; 30:4765-4774.