네이처 논문 번역

[Overview: 개요]

시험관 수정(IVF)은 자궁내막증, 난자의 질 저하, 부모의 유전 질환, 배란 문제, 정자 또는 난자에 해로운 항체 문제, 정자가 자궁경부 점액을 통과하거나 생존하지 못하는 문제, 낮은 정자 수 등으로 인해 발생하는 인간 불임 문제를 해결하는 데 널리 사용되는 방법이다. **그러나 IVF는 수정의 성공을 보장하지 않는다.** IVF를 선택하는 것은 높은 비용과 결과의 불확실성 때문에 부담이 크다. IVF 과정에서 발생하는 복잡성과 수정에 영향을 미치는 요인이 많아, 난임 전문의가 성공적인 출산 여부를 정확히 예측하는 것은 어려운 작업이다.

➔ IVF는 난임 문제를 위해 사용되는 방법이지만, 그 과정에서 발생하는 여러 요인으로 인해 반드시 성공을 보장하는 방법이 아니다.

본 연구에서는 인공지능(AI)을 활용하여 출산 성공 여부를 예측하였다. **특히, 기증자가 아닌 부부의 배아가 형성될 때 출산 성공 여부를 예측하는 데 초점을 맞추었다.** 본 연구에서는 Human Fertilisation and Embryology Authority(HFEA)가 제공하는 공개 데이터를 활용하여, 다양한 AI 알고리즘을 비교하였다. 전통적인 머신러닝 기법, 딥러닝 아키텍처, 그리고 여러 알고리즘을 조합한 앙상블 모델을 포함한 다양한 방법을 적용하였다. 데이터 분석 및 평가 지표로는 혼동 행렬(confusion matrix), F1-score, 정밀도(precision), 재현율(recall), 수신자 조작 특성(ROC) 곡선 등이 사용되었다.

모델 학습 과정에서는 **특징 선택 없이 학습(without feature selection) 하는 경우와 특징 선택 후 학습(with feature selection) 하는 경우 두 가지 설정으로 분류 모델을 학습시켰다.** 두 설정에서 머신러닝, 딥러닝, 앙상블 모델을 훈련하였으며, 특징 선택 없이 학습한 설정에서 랜덤 포레스트(Random Forest) 모델이 가장 높은 F1-score인 76.49%를 기록하였다. 동일한 모델에서 정밀도(precision), 재현율(recall), ROC AUC(곡선 아래 면적) 점수는 각각 77%, 76%, 84.60%로 나타났다.

➔ 인공지능을 활용하여 출산 성공 여부를 예측하였다. 이때 기증자에 대한 feature는 배제하였으며, feature selection과 모델 학습에 다양한 방식을 사용하였다.

임신의 성공 여부는 남성과 여성의 특성뿐만 아니라 **생활 환경에도 영향을 받는다.** 본 연구에서는 시험관 수정 과정에서 임상적으로 중요한 매개변수를 이용하여 임신 성공 가능성을 예측하였다. 이를 통해 인공지능이 진단, 예후 예측, 치료 결정 지원 등의 의사 결정 과정에서 유망한 역할을 수행할 수 있음을 보여준다.

➔ 임신의 성공 여부가 여러 요인의 영향을 받으므로 인공지능의 역할이 중요해질 것을 시사.

전 세계적으로 8천만 쌍 이상의 부부가 불임 문제를 겪고 있다. 약 12개월 동안 피임 없이 성관계를 가져도 임신이 성공하지 않는 경우 불임으로 간주될 수 있다. 이러한 임신 실패율을 줄이기 위해 여성의 난소에서 채취한 난자와 남성의 정자를 체외에서 수정시켜 배아를 생성한 후, 이를 여성의 자궁에 이식하여 임신을 유도하는 과정을 시험관 수정(IVF, In-Vitro Fertilization)이라고 한다. 일부 경우에는 인공 수정(Artificial Insemination)을 통해 정자를 직접 자궁 내로 주입하여 임신을 유도하기도 한다.

➔ 난임, IVF, 인공 수정에 대한 def.

현재까지 전 세계적으로 500만 명 이상의 아기가 IVF를 통해 태어난 것으로 보고되었다. IVF는 남성과 여성의 생식 기능과 관련된 다양한 문제로 인해 발생하는 불임을 극복하는 방법으로 활용된다. IVF 과정은 여러 의학적·외과적 절차를 결합하여 수정이 이루어지도록 돕는다. **IVF 치료는 한 번의 시술로 끝나는 것이 아니라 여러 차례의 주기를 거쳐 진행되며, 임신까지 몇 개월이 소요될 수 있다.**

IVF 치료의 접근성이 높아지면서, **불임이 아닌 부부들도 IVF를 선택하는 사례가 증가하고 있다. 그러나 IVF 치료는 높은 비용, 성공에 대한 불확실성, 치료 과정에서의 스트레스 등으로 인해 매우 도전적인 선택이 될 수 있다. 치료를 받는 환자들은 신체적·정신적 부담으로 인해 IVF 치료를 중단하는 경우가 많다.**

➔ IVF 시술의 여러 특성들 (시술의 복잡성, 증가하는 선호도, 과정에서 수반되는 스트레스)

여러 의료 전문가들은 임신 가능성을 예측하기 위해 경험에 기반한 시행착오 방법을 사용해 왔다. 따라서 기존의 예측 방법은 개별 의료 전문가의 경험 수준에 의존하며, 체계적인 통계적 접근법을 적용하지 않기 때문에 주관적인 요소가 강하다. 이에 따라 의료진과 환자들은 IVF 치료에 대한 의사 결정을 안내할 수 있는 보다 객관적인 측정 방법을 절실히 필요로 하고 있다.

최근 인공지능(AI), 머신러닝(ML), 딥러닝(DL)과 같은 기술의 발전은 데이터 기반의 통계적 접근법을 통해 이러한 문제를 해결하는 데 큰 가능성을 제공하고 있다. AI를 활용한 고도로 정확한 분석은 방대한 양의 데이터를 해석하여 의미 있는 결과를 도출함으로써, IVF 과정에서 발생하는 다양한 문제를 효과적으로 해결하는 데 기여할 수 있다.

이러한 통계적 접근법은 연구자들의 관심을 끌고 있으며, 이를 바탕으로 의료진이 IVF 과정에서 성공적인 출산 가능성을 보다 정확하게 예측할 수 있는 모델 개발이 활발히 진행되고 있다.

➔ 기존의 임신 가능성 예측에는 '경험적 방법'이 사용되었으며, 인공지능과 데이터 기반의 '통계적 접근'을 통해 문제에 대한 개선이 이루어지길 소망.

머신러닝(ML)은 컴퓨터 또는 시스템이 인간과 유사한 방식으로 사고하고, 과거의 경험을 학습하여 예측 결과를 출력하도록 하는 연구 분야이다. ML은 데이터에서 의미 있는 패턴을 탐색하며, 이를 통해 시스템이 인간의 의사결정 능력을 모방할 수 있도록 한다.

딥러닝(DL)은 머신러닝의 하위 분야로, 인간의 신경망(neural network) 원리를 기반으로 작동한다. 방대한 양의 데이터와 다양한 매개변수를 분석할 때, 인간은 중요한 패턴을 놓칠 가능성이 있지만, ML과 DL은 이러한 패턴을 효과적으로 찾아내어 보다 정확한 의사결정을 지원할 수 있다.

ML이 지속적으로 발전하면서, 여러 의료 분야에서 이미 의사결정을 향상시키기 위해 도입되고 있다. 대표적으로 맞춤형 의료, 수술 시뮬레이션, 신약 개발, 질병 진단 가속화 등의 분야에서 활용되고 있다.

생식과학 분야에서도 배아 이식 후 착상(implantation) 여부를 예측하기 위해 ML이 적용되었으며, 생존 출산(live birth) 예측 모델과 쌍둥이 출산 가능성을 평가하는 데에도 사용되었다. 최근에는 딥러닝 기법을 활용하여 치명적인 임신 관련 심장 질환 및 인간 배반포(human blastocyst) 선별을 예측하는 연구도 진행되고 있다.

따라서 ML은 임상 데이터를 활용하여 위험 평가, 진단 및 예후 예측 모델을 개발하고, 환자의 건강 관리 수준을 향상시키는 데 기여할 수 있다.

➔ ML/DL의 def와 의학 분야에서 이들의 중요도가 커짐을 강조.

과거 몇몇 연구에서는 머신러닝(ML) 기법을 활용하여 IVF 치료를 받는 여성의 출산 가능성을 예측하였다. 가장 초기이면서도 널리 받아들여진 예측 모델 중 하나는 McLernon 모델로, 이산 로지스틱 회귀(discrete logistic regression)만을 이용하여 최대 6회의 IVF 주기를 거친 부부의 출산 가능성을 예측한다.

이 연구에서는 두 가지 예측 모델이 개발되었다.

1. 사전 치료 모델(pre-treatment model): IVF 치료를 시작하기 전에 출산 가능성을 예측.
2. 사후 치료 모델(post-treatment model): 첫 번째 배아 이식 후 출산 가능성을 예측.

데이터는 1999년부터 2008년까지 IVF 치료를 시작한 253,417명의 여성 기록을 바탕으로 수집되었으며, 이들은 **자신의 난자와 배우자의 정자를 사용한 사례였다.** 연구 데이터는 영국 인간수정 및 배아생물학청(HFEA, Human Fertilisation and Embryology Authority)에서 제공되었다.

예측 모델의 성능 평가는 C-index(예측 일관성을 나타내는 지표)를 사용하여 측정되었으며,

1. 사전 치료 모델의 C-index: 0.69 (0.68–0.69)
2. 사후 치료 모델의 C-index: 0.76 (0.75–0.77)

즉, **사후 치료 모델이 더 높은 정확도를 보이며**, 첫 번째 배아 이식 후 출산 가능성을 예측하는 데 더욱 효과적이었다.

➔ 과거의 IVF에 대한 ML 모델 두 가지를 소개, 또한 이때 사후 치료 모델의 모델이 더 효과적이었음.

Rafiul Hassan 등은 IVF 임신 예측의 정확도를 높이기 위해 **힐 클라이밍(Hill-Climbing) 특징 선택 알고리즘과 다섯 가지 다른 머신러닝 모델을 활용한 연구를 제안하였다.**

이 연구의 데이터는 터키 이스탄불의 한 불임 클리닉에서 2005년 3월부터 2008년 1월까지 약 3년 동안 수집된 1,048명의 불임 치료 환자 기록을 포함하고 있다. 연구에서는 나이, 진단, 동난포 수(Antral Follicle Counts, AFC), 정자 질 등 총 27개의 속성을 사용하였다.

연구 결과, **임신 성공 여부에 가장 큰 영향을 미치는 IVF 속성은 나이**로 확인되었다. 또한, 힐 클라이밍 특징 선택 기법(중요한 특징만 선택하는 방법)을 적용했을 때, 모든 분류기의 성능이 향상되었다.

전체적으로 **서포트 벡터 머신(SVM)이 가장 높은 성능을 보였으며,**

1. 정확도: 98.38%
2. F1-score: 98.4%
3. AUC score: 99.5%

해당 모델은 IVF 관련 19개의 속성을 고려하여 최고의 성능을 달성하였다.

➔ Rafiul Hassan 팀의 연구 성과에 대한 소개 (힐 클라이밍 feature selection algorithm, 5 different ML Model and it's scores)

Guvenir et al.의 연구에서는 SVM, 의사결정나무(Decision Trees), 나이브 베이즈(Naïve Bayes), K-최근접 이웃(KNN) 등 여러 머신러닝(ML) 모델을 비교 분석하였다. 연구 결과, **모델마다 최적의 성능을 내기 위해 필요한 특징(feature) 수가 다르다는 점이 확인되었다.**

이 연구에서는 **환자의 나이, 체질량지수(BMI), 정자 수 등을 특징**으로 사용하여 모델을 훈련하였다.

1. SVM은 최대 64개의 특징을 사용하여 84%의 정확도를 기록하였다.

2. 반면, 인공신경망(ANN)을 사용한 Kaufmann et al.의 연구에서는 단 5~6개의 특징만 사용하였으며, 그 결과 59%의 정확도를 보였다.

이 연구에서는 두 가지 핵심 문제를 다루었다.

1. IVF 치료에서 임신 성공 가능성을 예측하는 문제

2. 의료진이 가장 생존 가능성이 높은 배아를 선택하도록 돕는 문제

➔ Guvenir et al. 팀의 경우 feature의 수를 조절하고, 그에 따른 학습 모델을 적용하여, feature의 개수와 학습 모델 사이의 상관관계의 중요성을 보여줌.

한편, Jiahui Qiu et al.의 연구에서는 **IVF 시행 전에 출산 가능성을 예측하는 모델을 개발**하였다.

이 연구에서는 로지스틱 회귀(Logistic Regression), 랜덤 포레스트(Random Forest), XGBoost(Extreme Gradient Boosting), SVM 총 4개의 모델을 비교하였다.

데이터는 2014년부터 2018년까지 중국 선징병원 의과대학(Medical Center of Shengjing Hospital of China Medical University)에서 첫 IVF 치료를 받은 7,188명의 여성 기록을 바탕으로 수집되었다.

연구에서 고려된 속성은 다음과 같다.

1. 나이, 항뮬러관 호르몬(AMH), BMI, 불임 기간, 이전 출산 경험, 이전 유산 경험 등

2. 불임 유형 (난관 폐쇄형, 남성 요인, 배란 장애, 원인 불명, 기타)

모델 성능 평가는 캘리브레이션(calibration)과 ROC 곡선(Receiver Operating Characteristic, ROC curves)을 활용하여 측정되었다.

그 결과,

1. XGBoost가 ROC AUC 0.73을 기록하며 가장 높은 성능을 보였으며,

2. 모든 모델 중에서 가장 뛰어난 캘리브레이션 성능을 나타냈다.

➔ Jiahui Qiu et al. 팀은 IVF 시행 전 출산 가능성 예측 모델을 개발함.

출산 성공 예측 문제는 이진 분류(binary classification) 문제에 속하며, 이는 주어진 IVF(시험관 아기) 관련 변수들을 기반으로 여성이 출산할지를 예측하는 것이다.

본 연구의 목표는 IVF 주기 완료 후 출산 성공 여부를 예측하는 다양한 모델을 비교하는 것이다.

이 연구는 기증자가 아닌 부부의 배아에서 발생하는 출산 성공 예측에 초점을 맞춘다.

*** 완전한 IVF 주기(Complete IVF Cycle)란?**

1. 신선 배아 이식(fresh cycle)
2. 동결-해동 배아 이식(freeze-thaw cycles)

위 과정이 하나의 난소 자극 주기(ovarian stimulation cycle)에서 이루어진 것을 의미한다.

*** 불임을 유발하는 생식적 요인**

불임은 여성과 남성 모두의 생식적 특징에 의해 발생한다.

*** 여성 관련 요인**

1. 나이 (난자의 수와 질 감소)
2. 생리 장애
3. 자궁 관련 요인
4. 자궁 경부 요인
5. 이전 임신 경험
6. 불임 기간
7. 원발성 불임(primary infertility): 최소 1년 이상 임신을 시도했으나 실패한 경우
8. 속발성 불임(secondary infertility): 이전에 임신 경험이 있으나 현재 임신이 불가능한 경우
9. 원인 불명의 불임

*** 남성 관련 요인**

1. 정자 농도
2. 정자 운동성
3. 정자 형태
4. 정액량
5. 총 정자 수

*** 데이터셋 개요**

본 연구에서는 위의 모든 주요 생식적 특징을 고려하여 분석을 진행하였다.

1. Human Fertilisation and Embryology Authority(HFEA)에서 제공하는 공개 데이터셋을 사용

2. 전 세계에서 가장 오랜 기간 IVF 치료 데이터를 축적한 데이터베이스

3. 2010년~2016년까지 영국 내 IVF 센터에서 수집된 495,630건의 기록

4. 총 94개의 임상적 특징 포함

5. 데이터 정제 후 141,160건의 기록을 최종 분석에 사용

6. 출산 성공(positive)과 실패(negative) 데이터 각각 70,580건씩 포함하여 균형 잡힌 데이터 구성

본 연구에서는 머신러닝(ML), 딥러닝(DL), 앙상블 학습(Ensemble Learning) 기법을 활용하였다.

*** 사용된 모델**

1. 머신러닝(ML) 모델
   - 로지스틱 회귀(Logistic Regression)
   - K-최근접 이웃(K-Nearest Neighbor, KNN)
   - 다층 퍼셉트론(Multi-Layer Perceptron, MLP)
   - 의사결정나무(Decision Tree)
2. 딥러닝(DL) 모델
   - 1차원 딥러닝 모델(1-D Deep Learning Model)
3. 앙상블 학습(Ensemble Learning) 모델
   - 랜덤 포레스트(Random Forest)
   - AdaBoost
   - 투표 분류기(Voting Classifier)

*** 특징 선택(Feature Selection) 기법**

본 연구에서는 두 가지 설정으로 모델을 학습하였다.

1. 특징 선택을 하지 않은 모델 학습 (without feature selection)
2. 특징 선택 기법을 적용한 모델 학습 (with feature selection)
   - Linear SVC(Linear Support Vector Classifier) 기반 특징 선택
   - 트리(Tree-based) 기반 특징 선택

*** 성능 평가 지표**

모든 모델의 성능은 다음과 같은 지표(metrics)를 활용하여 측정하였다.

1. F1-score
2. 정밀도(Precision)
3. 재현율(Recall)
4. ROC-AUC(Receiver Operating Characteristic - Area Under the Curve)

*** 논문 구성**

1. 방법론(Methodology): 데이터셋, 전처리 기법, 학습된 모델 설명
2. 결과 및 논의(Results & Discussion): 다양한 모델들의 성능 비교
3. 결론(Conclusion): 연구에서 얻은 인사이트 및 향후 연구 방향

[Methodology: 방법론]

1. 데이터셋 설명

본 연구에서 사용된 데이터셋은 Human Fertilisation & Embryology Authority(HFEA)에서 2010 년부터 2016 년까지 수집한 **익명화된 등록 데이터**이다.
이 데이터셋은 **세계에서 가장 오래된 불임 치료 관련 데이터베이스**로, 환자, 기증자, 자녀의 기밀 보호를 보장하면서 환자 치료 개선을 목표로 수집되었다.

데이터셋 세부사항

- **총 환자 기록**: 495,630 건
- **특징 수**: 94 개
- **수집 기간**: 2010 년~2016 년
- **데이터 유형**:
    - **숫자형 데이터**
    - **범주형 데이터**
    - **텍스트 데이터**

본 연구에서는 **부부의 행동 및 생리학적 특성에 대한 의료적 개입 없이** 데이터를 분석하였다. 또한 **부부 데이터를 분석**하는 것만을 대상으로 하였기 때문에 **기관윤리위원회(IRB)의 승인**은 필요하지 않았다. 모든 관련 지침은 이 연구에서 준수되었다.

목표 변수(출산 성공 여부)에 영향을 미치는 요인들

- **원본 데이터셋**에는 94 개의 특징이 포함되어 있지만, **모든 특징이 결과에 크게 영향을 미치는 것은 아니다**.

- 따라서, **30 개의 특징만 고려**하여 분석을 진행하였다.
- 특징 엔지니어링(Feature Engineering)은 **Bharti Bansal 박사**와 영국 국가보건임상우수지침(NICE)에서 추천한 주제 지식을 바탕으로 수행되었다.

선택된 특징

본 연구에서는 IVF 치료를 받는 여성의 **신선 주기(fresh cycles)** 및 이후 동결-해동 주기(freeze-thaw cycles)를 고려하였다.

- **기증자 난자/정자 주기**와 **PGD/PCS 주기**는 제외되었다.
- 포함된 주요 특징:
    - 나이
    - 이전 주기의 총 횟수
    - 이전 IVF 임신 횟수
    - 파트너 정자와 혼합된 난자 수
    - 이번 주기에서 이식된 배아 수
    - 불임 유형 및 원인 (남성 요인, 여성 요인, 배란 문제, 자궁내막증, 나팔관 문제, 자궁경부 문제 등)

표 1 은 연구에 사용된 데이터셋의 특징에 대한 자세한 설명을 요약한 표이다.

2. 데이터셋 전처리

원본 데이터셋에는 94 개의 특징이 포함되어 있지만, 그 중 일부는 **출산 성공 여부 예측**에 크게 영향을 미치지 않는다.
따라서 데이터셋의 필터링은 **자극 방법(stimulation used)**, **정자 출처(sperm source)**, 난자 출처(egg source)에 따라 이루어진다. 만약 정자와 난자의 출처가 동일한 부부(즉, 파트너와 환자)라면 해당 환자 기록만을 고려하고, 나머지는 제외된다. (* 즉 기증 data 에 대해서는 고려하지 않겠다는 뜻)

자극 방법

IVF 에서 여성에게 난포 자극 호르몬(FSH)과 황체 호르몬(LH)이 포함된 **주사제를 투여**하여 여러 개의 난자가 동시에 발달하도록 자극한다.
이는 데이터셋에서 "Stimulation Used"로 설명되며, 본 연구에서는 자극이 이루어진 환자 기록만을 고려한다.

나이 처리

"Patient Age at Treatment" 필드에서 몇몇 환자 기록에는 값 999 가 포함되어 있어 이를 제외한다. 또한, **텍스트와 나이 범위는 범주형 데이터로 변환**된다. 예를 들어, "Patient Age at Treatment" 필드에서:

1. 18–34 는 0 으로 변환
2. 35–37 은 1 로 변환
3. 38–39 은 2 로 변환

4. 40–42 은 3 으로 변환
5. 43–44 은 4 로 변환
6. 45–50 은 5 로 변환

출산 성공 여부 (Live-birth Occurrence)

"Live-birth Occurrence" 필드는 타겟 변수로, 값은 0 에서 5 사이로 존재하며, 0 은 출산이 없음을 의미하고(음성 클래스), 1 보다 큰 값은 출산이 있음을 의미한다(양성 클래스).
이 값을 **이진 분류**로 만들기 위해 **1 보다 큰 값**은 모두 1 로 설정하고, 나머지는 0 으로 설정한다.

데이터 불균형 문제 해결

위 필터링 후, **음성 샘플**이 **양성 샘플보다 5 배** 더 많아 데이터 불균형 문제가 발생한다. 이를 해결하기 위해 일부 음성 샘플을 제거하여 데이터 불균형을 조정한다.

최종 데이터셋

이제 최종 데이터셋에는 **141,160 개의 환자 기록**과 **25 개의 특징**이 포함되며, 각 클래스에 **70,580 개의 샘플**이 분포한다.
몇몇 필드(예: 정자 출처, 난자 출처, 불임 원인 등)는 **양성/음성 클래스에서 동일한 값을 가지며**, 분류에 중요한 영향을 주지 않으므로 이를 제거한다.

훈련 및 검증 세트 분할

샘플들은 훈련 세트(66%)와 검증 세트(34%)로 분할된다.

- **훈련 세트**: 93,165 개의 샘플
- **검증 세트**: 47,995 개의 샘플

데이터 정규화 및 상관 행렬

데이터는 **정규화(normalization)** 되어 있으며, 이는 데이터가 넓은 범위의 정수 값을 가질 수 있기 때문이다.
유사한 경향을 가진 특징들은 **상관 관계가 매우 유사할 가능성**이 있다. 이 경우, **하나의 특징만을** 모델에 공급하면 충분하다.
상관 행렬을 계산하여 각 특징의 중요도를 확인하고, 상관 계수가 **특정 임계값 이상인 특징들**은 **하나로 축소**된다.

(여기서부턴 무료 gpt 라서 번역 성능이 떨어짐.)

1. Linear SVC + SelectFromModel:

선형 모델은 L1 norm 으로 패널티를 부여받으면 희소한 해결책을 도출합니다: 많은 추정된 계수가 0 이 됩니다. 만약 목표가 데이터의 차원을 줄이고 다른 분류기를 사용하려는 경우, 이들은 scikit-learn 의 feature_selection. SelectFromModel 과 함께 사용되어 0 이 아닌 계수를 선택합니다. 희소 추정기는 이 목적에 유용합니다: 회귀를 위한 라쏘, 로지스틱 회귀, 그리고 Linear SVC. 이 방법에서 사용된 희소 추정기는 로지스틱 회귀, 결정 트리, 랜덤 포레스트, K 최근접 이웃 분류기입니다. 이 기법을 사용하면 특성 공간이 25 에서 20 으로 줄어듭니다.

2. Linear SVC + Tree-based feature selection:

랜덤 포레스트와 같은 트리 기반 추정기는 훈련이 완료된 후 각 특성의 중요도를 계산할 수 있으며, 이를 통해 특성 공간을 필터링하고 줄일 수 있습니다. 랜덤 포레스트에서는 훈련 중 각 특성에 대해 지니 불순도 또는 정보 이득/엔트로피를 사용하여 이 값을 계산하고, 이 방법으로 특성 중요도를 도출합니다. 특성 공간이 축소된 후, 우리는 이 새로운 세트를 사용하여 다른 추정기나 분류기로 훈련을 할 수 있습니다. 이 연구의 희소 추정기는 로지스틱 회귀, 결정 트리, 선형 판별 분석, 랜덤 포레스트, K 최근접 이웃입니다. 이 기법을 사용하면 특성 공간이 25 에서 5 로 줄어듭니다.

3. Deep learning: custom deep neural network:

ML 모델들과 함께, 동일한 데이터에 대해 딥 러닝 분류기(DL) 아키텍처도 훈련되었습니다. 이 신경망은 숫자 값을 (25 크기의 배열로) 입력으로 받기 때문에, 건축적 관점에서 1 차원 모델입니다. 출력층에는 시그모이드 활성화 함수를 가진 하나의 뉴런이 있어 이진 출력을 제공합니다(출산 발생 여부). 아키텍처는 총 9 개의 밀집(dense) 층을 포함하며, 각 층의 뉴런(모든 밀집 층에서)의 출력 값은 Rectified Linear Unit(ReLU) 활성화 함수를 통해 전달됩니다. 딥 러닝 분류기의 첫 번째 절반에서는 각 층의 뉴런 수가 이전 층의 두 배로 증가하며, 이 비율은 데이터셋의 성능 덕분에 균일하게 유지됩니다. 두 번째 절반에서는 층당 두 뉴런씩 감소하여 마지막 층에는 1 개의 뉴런만 남습니다. Adam 옵티마이저는 딥 러닝 신경망 훈련 중 손실 값을 최적화하는 데 사용됩니다. Adam 옵티마이저는 그라디언트가 희소하고 노이즈가 많은 문제에 적합한 최적화 알고리즘으로, AdaGrad 와 RMSProp 알고리즘의 장점을 결합하여 널리 사용되고 있습니다. 이 연구에서는 Adam 옵티마이저의 성능이 다른 옵티마이저들보다 우수하다는 결과를 확인했습니다.
딥 러닝 분류기를 훈련시키는 총 에포크(epoch) 수는 50 입니다. 이 데이터셋에서 과적합을 방지하기 위해 **Dropout** 과 **배치 정규화(Batch Normalization)**, **조기 중지(Early stopping)** 기법이 사용되었습니다. 뉴런 수가 증가할수록 딥 러닝 분류기의 중간 밀집 층에서 노이즈 생성 확률이 높아지기 때문에, 중간 밀집 층(512 유닛) 이후에는 20%의 드롭아웃을 도입했습니다. 이진 분류 설정에 적합한 손실 함수인 **이진 교차 엔트로피(Binary cross-entropy loss)**가 사용됩니다. 전체 데이터셋에 대해 그라디언트를 계산하는 것이 비용이 크므로, 에포크당 128 개의 샘플을 배치 크기로 하여 훈련을 진행했습니다. 커스텀 딥 러닝 아키텍처의 개요는 그림 3 에 나와 있습니다.

앙상블 학습. 앙상블 방법은 우수한 성능으로 인해 머신러닝(Machine Learning) 분야에서 가장 유명한 학습 알고리즘입니다. 이 방법의 핵심은 여러 개의 머신러닝 알고리즘을 결합하여 정확한 결정을 내린다는 것입니다. 본 연구에서는 다음의 알고리즘들이 학습되었습니다.

1. 랜덤 포레스트 (Random Forest)
2. 아다부스트 (AdaBoost)
3. 투표 분류기—소프트/하드 (Voting classifier—soft/hard)

투표 분류기 (Voting Classifier)

투표 분류기는 여러 분류 모델들의 결과를 결합하여 최종 결정을 내리는 방법입니다. 예를 들어, 5 개의 이진 분류 모델이 학습되어 미지의 샘플을 예측한다고 할 때, 각 모델의 예측값을 투표 시스템에 입력하여 최종 예측을 도출합니다. 보충 그림 1 은 투표 분류기를 설명합니다.

투표 시스템은 두 가지 전략을 따릅니다: **하드 투표 (Hard Voting)**와 **소프트 투표 (Soft Voting)**.

- **하드 투표**는 다수결 투표 방식으로, 개별 모델들 중 가장 많은 표를 얻은 클래스가 선택됩니다. 만약 Nc 가 클래스의 투표 수이고, y1, y2, y3, ..., yn 이 n 개의 다른 분류기들의 예측이라면, 하드 투표 공식은 Eq. (1)로 표현됩니다.
- **소프트 투표**는 각 분류기의 확률 점수 벡터를 입력으로 받아 이를 합산하고, 그 후 평균을 구합니다. 최종 출력 클래스는 가장 높은 확률 점수를 얻은 클래스가 됩니다. 만약 p1, p2, ..., pn 이 n 개의 다른 분류기들의 확률 점수라면, 소프트 투표 공식은 Eq. (2)로 표현됩니다.

이 투표 분류기에 사용된 분류기는 **로지스틱 회귀 (Logistic Regression)**, **결정 트리 (Decision Tree)**, **선형 판별 분석 (Linear Discriminant Analysis)**, **랜덤 포레스트 (Random Forest)**, **K-최근접 이웃 분류기 (K Nearest Neighbours)**입니다. 다음 섹션에서는 이 모델을 실험 데이터셋을 통해 검증하였습니다.

[결과 및 논의]

이 연구에서는 딥러닝 분류기 훈련을 위해 Keras 백엔드를 사용하는 TensorFlow 라이브러리와 머신러닝 분류기를 위해 scikit-learn 을 사용했습니다. 본 연구에서 비교된 메트릭은 F1-스코어, 정밀도(Precision), 재현율(Recall), ROC AUC 점수와 곡선입니다. 이 섹션에서는 Feature Selection 을 적용한 모델과 적용하지 않은 모델의 결과를 제시합니다. 비교 테이블에는 ML 기반, DL 기반, 앙상블 기반 모델 등 주요 범주가 표시됩니다. 표 2 는 Feature Selection 을 적용하지 않은 모델들 간의 성능 비교를 설명합니다.

Feature Selection 을 적용하지 않은 설정 결과

표 2 는 앙상블 학습 모델이 재현율, F1-스코어, ROC AUC 점수에서 더 나은 분류 성능을 보였음을 나타냅니다. Random Forest 는 F1-스코어에서 76.49%로 가장 높은 점수를 기록했습니다. Random Forest 의 재현율 값은 다른 모델들에 비해 눈에 띄게 높았으며, 76%를 기록했습니다. 그림 4a 는 Feature Selection 을 적용하지 않은 모델들에 대해 훈련된 ROC AUC 곡선을 나타냅니다. Random Forest 는 84.6%로 가장 높은 AUC 점수를 기록했습니다.

Feature Selection 을 적용한 설정 결과

방법: Linear SVC + SelectFromModel. 표 3 은 앙상블 학습 모델이 더 나은 분류 성능을 보였음을 설명합니다. Multi-Layer Perceptron 과 AdaBoost 가 F1-스코어에서 72.98%로 가장 높은 점수를 기록했습니다. AdaBoost 는 77.60%로 가장 높은 ROC AUC 점수를 달성했습니다. 그림 4b 는 Feature Selection 을 적용한 설정에서 훈련된 모델들의 ROC AUC 곡선을 나타냅니다. 이 방법에서 AdaBoost 는 77.60%로 가장 높은 AUC 점수를 기록했습니다. 이전 결과와 비교한 표 2 를 보면, Feature Selection 방법이 ROC AUC 점수, F1-스코어, 재현율 등 메트릭에서 전체 성능을 감소시키는 영향을 미쳤음을 알 수 있습니다.

방법: Linear SVC + Tree-based feature selection. 이 방법은 Extra Trees Classifier 를 사용하여 트리 기반 특성 추출기를 적용합니다. Extra Trees Classifier 는 랜덤 포레스트와 약간 다릅니다. 그 차이는 부모 노드를 두 개의 랜덤 자식 노드로 나누기 위해 랜덤하게 분할을 선택한다는 점입니다.

표 4 는 다시 머신러닝 기반 분류 모델이 더 나은 성능을 보임을 나타냅니다. 이 특성 선택 방법에서 얻은 가장 높은 F1-스코어는 73.46%로, 이전 방법보다 낮습니다. 최대 재현율 값은 72%로, 이전 방법과 동일합니다. 그러나 ROC AUC 점수는 Deep Learning 분류기와 AdaBoost 를 제외하고 이전 방법보다 증가했습니다. 그림 4c 는 Linear SVC + Extra Tree classifier 로 훈련된 모델들의 ROC AUC 곡선을 나타냅니다. AdaBoost, MLP, DL 분류기들이 가장 높은 AUC 점수를 기록했습니다.

위의 세 가지 방법을 비교한 결과, 특성 선택 방법을 사용하지 않은 일반적인 특징들, 특히 랜덤 포레스트(앙상블 학습 방법)가 76.49%의 더 나은 정확도와 84.6%의 AUC 점수를 기록한 것이 명확하게 더 나은 성능을 보였습니다. 따라서 실시간 결과를 위한 프로덕션 환경에서는 이 모델을 사용하는 것이 바람직합니다.

[결론]

클리닉에서는 의료 제공자들이 경험이나 자신이 속한 난임 치료 센터의 성공률을 바탕으로 생아 출산에 대한 상담을 제공할 수 있지만, 이는 일부 경우에는 부적절할 수 있습니다. 본 연구는 환자의 자연적으로 측정 가능한 예측 변수를 바탕으로 성공적인 또는 실패한 IVF 치료를 예측하는 도구가 환자와 의료 제공자에게 구체적인 결정을 내릴 수 있도록 돕는 데 기여할 것입니다. 이 도구는 IVF 치료를 받기 전에 비용이 많이 들고 번거로운 과정을 겪기 전에 생아 출산 가능성에 대해 부부에게 상담을 제공할 것입니다. 이 도구는 다른 모델들과 비교해 84.60%의 AUC 점수와 76.49%의 F1-스코어를 기록했습니다. 그러나 현재로서는 이 도구만으로 의사 결정을 의존하는 것은 권장되지 않습니다. 왜냐하면 데이터가 단일 출처에서 수집되었기 때문에 모든 인구에 일반화할 수 없기 때문입니다. 모델들은 제한된 요소들에 대해 훈련되었으며, 알콜 소비, 흡연, 카페인 소비, 고혈압 및 다른 생활 습관과 같이 임신 예측에 중요한 영향을 미치는 여러 요소들은 데이터셋의 제한으로 인해 고려되지 않았습니다.

미래 연구의 범위는 다양한 지리적 위치에 있는 IVF 클리닉들에서 데이터를 수집하여 전 세계 여러 인종에 대한 정보를 포함시키는 것입니다. 개인의 생활 습관에 관한 몇 가지 파라미터를 고려해야 할 필요가 있으며, 이러한 세부사항들은 간접적으로 난임에 영향을 미칩니다. 다양한 인종과 연령대의 데이터를 수집하면 AI 성능이 개선될 수 있습니다. 또한 IVF 에서 각 특징의 중요성을 강조하는 성공적인 난임 연구를 진행할 수 있습니다. 모델 성능을 향상시키기 위해 다른 특성 선택 및 차원 축소 방법을 사용할 수 있습니다.