후성유전 분야에서의 인공지능 활용현황과 미래 방향성

연세대학교 대학원
의료법윤리학협동과정
보건학전공
박 지 원

후성유전 분야에서의 인공지능 활용현황과 미래 방향성

지도교수 김 소 윤

이 논문을 석사 학위논문으로 제출함

2024년 1월

연세대학교 대학원

의료법윤리학협동과정

보 건 학 전 공

박 지 원

박지원의 석사 학위논문을 인준함

심사위원 김 소 윤 인

심사위원 김 태 헌 인

심사위원 이 유 리 인

연세대학교 대학원

2024년 1월

감사의 글

석사과정의 마침표인 학위논문을 마무리한 지금, 이 순간의 감정과 지나온 시간들은 저에게 앞으로 나아갈 자양분과 든든한 밑거름이 될 것으로 믿어 의심치 않습니다.

인턴연구원 시절부터 석사과정을 마치는 지금까지 약 3년이란 시간 동안 지도해주시고, 거시적인 관점에서 미래를 바라볼 수 있도록 일깨워주신 김소윤 교수님께 깊은 감사의 말씀 올립니다. 석사과정으로 한 발 나아가도록 응원해주시고 더 넓은 사회와 다양한 연구를 경험하며 성장하게 해주신 이유리 교수님께도 늘 감사드립니다. 갑작스럽게 부심을 요청드렸음에도 귀중한 시간 내시어 나무가 아닌 숲의 관점에서 바라볼 수 있도록 지도해주시고, 따뜻한 말씀과 더불어 날카로운 조언을 주신 김태현 교수님께도 감사드립니다. 이외에도 항상 유쾌하지만 진중하게 가르쳐주시고 고민 들어주시는 이동현 교수님, 국제보건 분야에서 많은 연구를 통해 성장할 수 있도록 가르침 주신 용태순 교수님, 안동일 교수님, 윤문수 교수님, 그리고 강선주 교수님께 감사드립니다. 깊은 연륜과 경험을 바탕으로 어디서도 배울 수 없었던 가르침을 주신 김강립 교수님, 학부시절부터 지금까지 늘 함께 고민해주시고 격려해주시는 이준영 교수님께도 특히 감사드립니다.

연구원 생활에 있어 많은 도움과 힘이 되어주신 의료법윤리학연구원 박사님들과 선생님들, 305호 조교실 선생님들께도 감사합니다. 특히 입학 동기로서 만나 조교실에서 함께 생활하며 기쁨과 슬픔을 공유할 수 있었던 소중한 김소연 연구원에게 감사합니다. 마지막으로 어떤 선택과 결정을 하든, 항상 묵묵히 응원해주고 지지해주는 사랑하는 부모님께 더할 나위 없이 감사드립니다. 하나뿐인 남동생과 끊임없는 사랑과 지지를 보내주는 소중한 인연에도 고맙다는 인사를 전합니다. 감사합니다.

2024년 1월, 박지원 올림

차 례

표 차례

그림 차례

국문요약

후성유전 분야에서의 인공지능 활용현황과 미래 방향성

인공지능 기술은 복잡하고 많은 양의 데이터를 통합, 해석 및 관리하는 데 유용하며 연구자와 임상 현장에서의 의사결정에 중요한 역할을 한다. 특히 후성유전 분야에서 DNA 메틸화 패턴을 결정하기 위한 예측모델 등과 같은 다양한 인공지능 기술의 적용은 후성유전학적 연구를 용이하게 하며, 개인 맞춤형 정밀의료를 제공할 수 있는 중요한 요소로 자리매김하고 있다.

이런 측면에서 본 연구는 후성유전 분야에서 인공지능이 어느 정도 범위까지 적용되고 있는지 확인하고자 하였다. 주제범위 문헌고찰 결과 후성유전과 관련하여 인공지능을 구현한 연구에서는 질병의 발생 예측, 환자군 분류, 치료의 차별적 예후를 확인하기 위한 목적으로 인공지능 기술이 적용되고 있었다. 추가로 전통적 문헌고찰 방법을 활용하여 역학·유전학·후성유전학 분야에서 인공지능 기술을 어느 정도 범위까지 적용하고 있는지 탐색 및 비교하였다. 전통적 문헌고찰 결과, 역학 및 유전학 분야에서는 챗봇 등 추론 결과를 기반으로 물리적 행동 수행이나 시스템 처리를 자동으로 유발하는 '행동' 영역의 인공지능 모델이 활발히 활용되고 있었으나, 후성유전 분야에서는 해당 영역의 적용이 아직 이뤄지지 않고 있다는 점을 확인할 수 있었다.

종합적인 결과를 바탕으로 '데이터 통합성 향상을 위한 결합형 빅데이터 활용 활성화'와 '후성유전 분야에서 인공지능 기술범위의 확대 필요성'을 후성유전 분야에서의 인공지능 적용 미래 방향성으로 제시하였다. 제시된 미래 방향성은 개인에게 적합한 생활습관 관리, 예방적 관점의 디지털 헬스케어를 가능하게 할 것으로 기대

되며, 우리나라의 보건의료빅데이터플랫폼 활성화 및 국가 바이오 빅데이터 플랫폼의 필요성을 뒷받침하는 참고자료로 활용될 수 있을 것으로 기대된다.

그러나 이에 앞서 빅데이터 플랫폼에 포함하기 위한 후성유전 정보 수집 및 활용의 범위를 설정하는 것이 필요하며, 정보 수집으로 인한 차별 문제를 예방하기 위해 법·윤리적 규제 마련에 대한 사회적 논의가 선행되어야 할 것이다. 이러한 논의가 선행되었을 때, 본 연구를 통해 개인 맞춤형 정밀의료를 가능하게 할 수 있을 것이다. 궁극적으로 개인 수준의 건강관리에서 나아가 집단 수준의 건강관리까지 가능하게 하는 공중보건의 목표 달성을 가능하게 할 수 있을 것이다.

핵심어 : 후성유전, 정밀의료, 인공지능, 기계학습, 미래 방향성, 유전정보, 빅데이터, 생물학적 데이터

제1장 서론

1.1. 연구의 배경 및 필요성

2003년 인간게놈프로젝트 이후, 유전체 의학은 많은 발전을 이루어 왔다(HapMap Consortium, T.I., 2003). 인간게놈프로젝트를 통해 발견하게 된 유전체 정보를 활용하여 개인의 유전적 소인에 따라 진단과 치료를 제공하는 맞춤의학이 도래하였다(류제운 등, 2013). '맞춤의학(Personalized Medicine)'은 보건학적 통계에 근거한 표준 치료방법과 달리 가족력, 위험인자, 고유병력 등 개인특성을 확인하고 유전적 특성인 '유전자형(Genotype)'과 '유전자 내 발현 프로파일(Gene Expression Profile)' 등의 차이를 고려하는 치료방법이다(Abrahams, 2005). 이를 통해 질병 발생의 효과적 예측, 치료의 효율성 확대, 부작용의 최소화가 가능하며, 환자의 만족 증대, 의료비용의 효율적 운용이 가능하다.

현대의학은 치료중심이 아닌 예방중심이라는 특징을 가지며 이러한 측면에서 '정밀의학(Precision Medicine)'은 미래 의학의 주된 이슈로 손꼽힌다. 정밀의학은 임상병리학에 분자의학 기술을 도입하여, 유전·환경·생물학적 특성 등 환자 개인에게 적합한 진단 및 치료를 한다는 개념이다. 맞춤형 의학은 정밀의학의 기반 중 하나이며, 개인의 고유한 특성에 의료를 적용하는 것을 목표로 한다. 이를 위한 노력으로 질병 감수성과 약물 반응을 지배하는 유전적 변이의 역할을 명확하게 하는 데 집중되었다(Stower, 2020; Delhalle et al., 2018).

그러나 유전체 서열을 분석하는 것만으로 확인이 어려운 경우가 있다. 후천적 요인인 환경이나 식이, 운동 등 생활습관 또한 유전적 요인에 못지않게 개인의

건강에 영향을 준다. 이를 유전자의 후성유전학적 변화라고 한다. 후성유전학적 변화는 인간의 건강 전반에 걸쳐 영향을 미친다(Jaenisch & Bird, 2003; Venter et al., 2001). 이러한 변화를 연구하는 생물학의 한 분야를 후성유전학(Epigenetics)이라고 한다(Feinberg, 2018). DNA 염기서열의 변화는 발생하지 않으나, 유전자 발현에 있어 변화가 발생하는 것에 관한 연구로 정의된다. 개인의 유전자형·나이·식생활·알코올 소비·흡연과 같은 생활습관 사이의 복잡한 상호작용의 산물로 가정된다(Rauschert et al., 2020).

유전자와 환경 간의 상호작용은 인간이 환경에 노출되는 정도에 따라 상이하며 환경뿐 아니라 생활습관과도 관련되어 있다. 이러한 상호작용은 태어날 때 부모로부터 물려받은 DNA와 유전자에 변형을 유발하고, 질병 발현의 양상을 달라지게 한다(강길전, 2010; 오정환 등, 2008). 즉 후성유전학적 환경에 관한 연구는 유전체 서열 분석으로 확인하기 어려운 정보를 제공할 수 있다. 후성유전학은 초기 생애의 부정적 환경과 후기 질병 발병 간의 중재 역할을 하는 것으로 확인되고 있어, 유전학만큼이나 질병 진단 및 치료에 적합한 학문이라 할 수 있다(Rauscher et al., 2020).

이에 후성유전적 변화와 관련된 후성유전, 생물학, 질병 이해에 대한 예측이 매우 중요하다. 후성유전체의 변화가 질병의 원인인지, 질병의 결과로 후성유전체가 변화하는 것인지에 대한 상관관계의 확인을 위해 생물학적 통계 도구, 실험연구, 생화학적 연구 등이 도구로서 요구된다(Feinberg, 2018). 후성유전 정보는 생물학적 데이터의 특성을 가지며, 많은 양의 분자를 가지고 있기에 이를 분석하는 고처리 절차가 요구된다. 이러한 특성을 가진 데이터를 'Omics 데이터'라고 한다. Omics 데이터는 변수와 잡음이 많고, 샘플 수가 상대적으로 적어 희소하다(Xu & Jackson, 2019).

이에 인공지능 기술을 적용함으로써 발생 가능한 자원의 낭비를 최소화할 수 있다(Hawkins et al., 2023). 인공지능 기술은 복잡하고 많은 양의 데이터 세트를 통합, 해석 및 관리하는데 유용하며, 연구자와 임상 현장에서의 의사결정 혹은 우선순위 결정을 지원하는 데 중요한 역할을 할 수 있다(Rauschert et al., 2020; Aroa & Tollefsbol, 2021). 특히 후성유전 분야에서 DNA 메틸화 패턴을 결정하기 위한 예측모델 등과 같은 다양한 인공지능 기술의 적용은 후성유전학적 연구를 용이하게 한다(Rauschert et al., 2020; Holder et al.,2017). 인공지능 기술은 후성유전 측면에서 효과적인 방법으로 제시되고 있으며 후성유전학적 정보에 기초하여 개인 맞춤형 치료를 제공할 수 있는 중요한 요소로 자리매김하고 있다(De Riso & Cocozza, 2021; Holder et al., 2017; 김휘영, 2018).

후성유전 분야의 인공지능 적용과 관련하여 암, 심혈관 질환과 같이 특정 질병 분야에 인공지능의 적용현황을 분석한 연구와 후성유전 분야 전반에 걸쳐 인공지능 기술의 적용을 고찰한 연구는 확인 가능하였다. 그러나 인공지능 기술적용 수준 고찰에서 더 나아가, 미래 방향성을 제안한 연구는 확인할 수 없었다.

이에 본 연구에서는 후성유전 분야에서 인공지능 기술의 적용현황을 파악하고, 후성유전 분야에서의 인공지능 기술적용 미래 방향성을 제시해보고자 한다.

1.2. 연구의 목적

후성유전 분야에서 현재 인공지능 기술이 어느 정도 영역까지 활용되고 있는지 현황을 파악하고, 정밀의료의 목표를 달성하고 예방적 관점의 건강관리를 위한 인공지능 적용의 미래 방향성을 제시하기 위함이다.

구체적인 연구목적은 다음과 같다.

첫째, 후성유전 분야에서의 인공지능 기술적용 현황과 활용 데이터, 적용 목적을 확인한다.

둘째, 역학 · 유전학 · 후성유전 분야에서의 인공지능 기술적용 영역을 탐색하고 비교한다.

셋째, 후성유전 분야에서의 인공지능 기술적용 미래 방향성을 제시한다.

1.3. 연구의 방법

본 연구는 후성유전 분야에서 인공지능의 적용에 관한 현황을 파악하고자 서술적(descriptive) 설계로 진행한 문헌고찰 연구이다. “후성유전 분야에 인공지능 기술은 어느 정도 적용되어 있는가?” 라는 연구 질문으로 시작한 문제의 답을 찾기 위해 주제범위 문헌고찰(Scoping Review) 방법을 사용하였다.

주제범위 문헌고찰은 잘 알려지지 않은 특정 주제의 연구를 개괄하고 정책이나 연구를 위한 이용 가능한 근거를 요약하는 데 유용하며, 연구설계나 존재하는 근거의 유형을 빠르게 검토하는 것을 목적으로 한다(천희란 등, 2022). Arksey & O’Malley(2005)에 따르면 주제범위 문헌고찰 연구의 목적과 유형은 다음과 같이 구분해볼 수 있다. 첫째, 주제 관련 연구의 범위와 속성을 빠르게 파악하는 것(mapping the fields)이다. 둘째, 체계적 문헌고찰이 필요한지 파악하기 위해 선행연구 단계로 수행하는 것이다. 셋째, 관련 주제의 연구결과를 요약하고 확산하는 것이다. 넷째, 선행연구가 없는 영역을 찾고 현존 문헌에서 지식의 틈을 좁히는 것 등이다(천희란 등, 2022).

본 연구에서는 첫 번째 유형인 ‘주제 관련 연구의 범위와 속성을 빠르게 파악하는 것(mapping the fields)’ 을 목적으로 하였다. Arksey와 O’Malley(2005)의 방법론적 틀에 따라 1) 연구 질문 개발 2) 관련 연구 검색 3) 문헌 기입과 정리 4) 자료 요약 5) 결과 보고의 연구절차에 따라 수행하고자 하였다.

추가로 전통적 문헌고찰 방법(Traditional Literature Review)을 활용하여 후성유전 분야에서의 인공지능 적용 한계점을 파악하고 미래 방향성을 제시하고자 하였다. 문헌고찰 방법은 현존하는 연구 및 자료를 분석하여 현재 동향을 파악하고, 기술의 발전 방향, 잠재적 문제점 및 해결 방법, 시장 및 산업 내에서의 사

용 사례 등을 파악하는 연구방법이다. 선행연구 및 기타 자료를 중심으로 역학·유전학·후성유전 각 분야에서 인공지능 기술을 어느 정도 영역까지 적용하고 있는지 탐색 및 비교하였다. 비교 결과를 바탕으로 향후 후성유전 분야에서 인공지능이 어떻게 적용되고 발전되어야 할 것인지에 대한 방향성을 제안하고자 하였다 (그림 1).

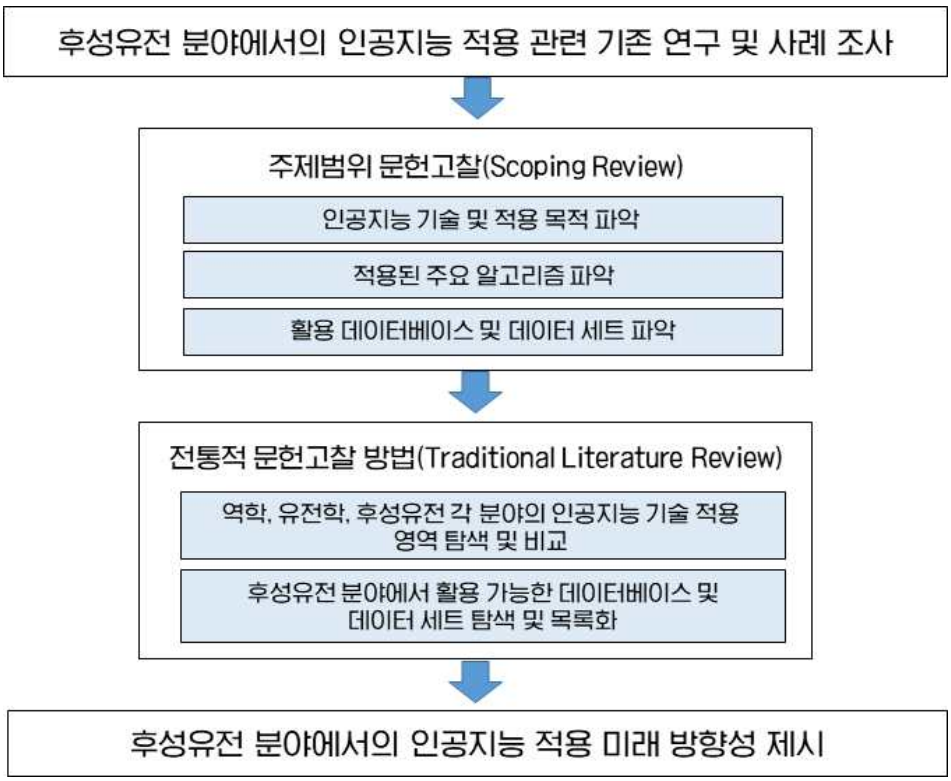

그림 1. 연구수행 흐름도

1.3.1. 연구 질문 개발

이 연구의 목표를 달성하기 위해 다음과 같이 세 가지 연구 질문을 개발하였다.

첫째, 후성유전 분야에서 실제 인공지능 기술을 구현한 연구의 특징은 무엇인가?
둘째, 후성유전 분야에서 실제 구현된 인공지능 기술과 적용 목적, 활용 데이터 베이스 및 데이터 세트는 무엇인가?
셋째, 역학·유전학·후성유전 각 분야에서 인공지능 기술이 어느 정도 영역까지 적용되어 있으며 후성유전 분야에서의 적용 한계점은 무엇인가?

1.3.2. 관련 연구 검색

'후성유전 분야에서의 인공지능 적용현황' 을 검토하기 위하여 국내·외 각종 학술지 게재논문을 대상으로 현황을 분석하였다.

첫째, 국내의 경우 연세대학교 학술정보원의 검색서비스를 통해 학술연구정보서비스(RISS), 한국학술정보(KISS)를 활용하였다. 국외의 경우 Pubmed를 활용하였다. Pubmed는 Medline을 포함한 대부분의 의학분야 논문을 확인할 수 있으므로, 기타 검색서비스를 통한 검토는 배제하였다. 또한, Google Scholar에서 주요어 'epigenetic', 'Artificial Intelligence' 를 이용하여 수기 검색을 보완적으로 수행하였다.
둘째, 검색 키워드는 선행연구를 참고하여 설정하였으며, 통제어와 AND, OR 및 절단검색 기능을 조합하여 검색하였다(표 1).

표 1. 검색 키워드 및 조합

구분		인공지능 관련	후성유전 관련
키워드	국문	인공지능, 컴퓨터 지능, 기계학습, 컴퓨터 보조 러닝, 머신러닝, 딥러닝	후성유전, 후생유전
	영문	Artificial Intelligence, Computational intelligence, machine intelligence, computer assisted learning, machine learning, deep learning	Epigenome, Epigenomic, Epigenetic, Episignature, Epivariation, Epimutation, Methylome, Methylomi, DNA methylation, Histone modification, Noncoding RNA, Non-coding RNA, ncRNA, microRNA, miRNA, Chromatin
검색 조합	국문	(인공지능 OR 컴퓨터지능 OR 기계학습 OR 컴퓨터 보조 러닝 OR 머신러닝 OR 딥러닝) AND (후성유전 OR 후생유전) AND (정보 OR 데이터)	
	영문	(Artificial Intelligence OR Computational intelligence OR machine intelligence OR computer assisted learning OR machine learning OR deep learning) AND (Epigenome OR epigenomic OR epigenetic OR episignature OR epivariation OR epimutation OR methylome OR methylomic OR DNA methylation OR Histone modification OR Noncoding RNA OR ncRNA OR microRNA OR miRNA OR Chromatin)	

참고: Brasil, et al., (2021). Artificial intelligence in epigenetic studies: Shedding light on rare diseases. *Frontiers in Molecular Biosciences,* 8, 648012.

1.3.3. 선정 및 제외기준

본 연구는 2013년부터 2023년까지 후성유전 분야 인공지능 적용과 관련하여 출간된 국내·외 학술지 게재논문 전수를 포함하였다.

첫째, (출판물의 종류) 학술 수준의 정보를 포함하기 위해 보고서, 도서, 학위논문을 제외한 학술논문(Article)만을 확인하고자 하였다.

둘째, (발행 연도) 우리나라의 경우 공공기관이 보유·관리하는 데이터의 제공 및 그 이용 활성화에 관한 사항을 규정함으로써 공공데이터 이용권을 보장하고, 민간 활용을 통한 삶의 질 향상과 국민경제 발전에 이바지함을 목적으로 2013년 「공공데이터 제공 및 이용 활성화에 관한 법률」을 제정 및 시행, 고수요·고가치·대용량의 36대 주요 데이터를 선정 및 개방하였다(공공데이터법, 2013). 해당 법 시행 이후 공공데이터 활용 관련 연구가 대폭 확대된 바 있어, 앞서 설정한 국문 검색 키워드(정보 또는 데이터) 관련 연구를 최대한 포함하고자 하였다(정보영, 2018). 또한 후성유전 분야의 인공지능 적용은 계속 발전되고 있음에 따라 현재 적용현황과 부합하지 않는 과거의 문헌을 배제하고자 하였다. 이에 2013년부터 2023년까지 최근 10년 이내에 발간된 논문으로 게재연도를 제한하였다.

셋째, (작성언어) 연구 논문의 공식 언어인 영어를 채택함으로써 포함되는 논문의 수를 최대화하고자 하였다. 한국어의 경우 국내 학술 데이터베이스를 활용하기 위해 포함하였다.

넷째, (기타) 선행연구(Sharma et al., 2022)를 참고하여 선정·제외 기준을 수립하였다. 실제 적용을 확인하는 것이 목적이므로 개념증명, 타당성 조사, 제언과 같은 연구는 제외하였다. 질병의 후성유전적 발생기전 분석을 위해 인공지능을 적용한 양적연구를 포함하였다. 2013년 이전 출판, 전문 미제공, 인공지능의 개념 및 유형만을 정리했거나, 인공지능을 적용하지 않은 연구를 제외하였다. 질병의 발생 요인에 관해 인공지능을 적용하였으나 후성유전과 관련된 내용을 다루지 않은 연구, 질병의 발생 등을 확인한 역학적 연구, 사람이 아닌 다른 종(예. 식물)을 대상으로 한 연구 등의 경우 제외하였다.

제2장 후성유전과 인공지능

2.1. 후성유전의 개념

2.1.1. 후성유전의 정의

‘후성유전학’은 ‘DNA 서열의 변화로 설명할 수 없는 유전자 발현의 유산’ 또는 ‘특정 메커니즘에 의한 유형 또는 유전자 발현의 변화’라는 전통적 정의를 가진다. 일반적 정의는 ‘유전적 서열의 수정을 포함하지 않는 유전 기능의 가역적인 변화’이다. 세포가 정보를 켜고 끌 수 있다는 개념은 Spemann & Mangold(1924)에 의해 소개되었으며, ‘후성유전학(Epigenetics)’이라는 단어는 1942년 Conrad H. Waddington에 의해 처음 언급되었다. Waddington의 초기 정의에 따르면, 후성유전학은 ‘유전자와 생성물 간 상호작용을 연구하는 생물학의 한 분야로서 표현형의 생성과정을 연구하는 학문’이다(Waddington, 1942).

후성유전학은 Waddington에 의해 정의된 이후 여러 차례 재정의되었다(Cavalli & Heard, 2019). Nanney(1958)는 RNA 또는 DNA 서열을 의미하는 유전물질의 발현을 조절하는 시스템이라고 정의하였다. Riggs et al(1996)은 DNA 서열의 변화로 설명할 수 없는 유전자 기능의 유사분열 또는 유사분열적으로 유전될 수 있는 변화라고 정의하였다. Bird(2007)는 염색체 영역의 구조적 적응으로, 변경된 활동 상태를 등록·신호 또는 지속시키는 역할이라 정의하였다. Lappalainen & Greally(2017)는 유전체 조절자를 통해 매개되는 세포의 특성으로, 세포가 과거 사건을 기억할 수 있는 능력을 부여하는 것이라 하였다. Nicoglou(2017)는 유전자 잠재성에 대한 작용을 통해 발달과정의 안정성에 영향을 미치는 다양한 세포 내 인자를 후성유전이라고 정의하였다.

후성유전학을 뜻하는 Epigenetics의 'Epi' 는 그리스어로 'on' 또는 'over' 라는 의미를 가진다. 유전자 구조가 아닌 유전자 구조의 외부에서 유전자 발현이 조절된다는 것이다(강길전, 2010). 세포가 분열되는 동안 DNA 염기구조 또는 크로마틴의 변형을 통하여 유전자의 발현 양상이 변하고, 표현형이 변화하게 된다(배다정 & 박춘식, 2013; Yang & Schwartz, 2011). 변화는 부분적으로 유전적이며, 환경과의 상호작용에 의해 결정된다(De Riso & Cocozza, 2021). 이러한 사례는 때때로 일란성 쌍둥이에게서 확인할 수 있다. 유전적으로 동일한 DNA를 가졌으며 같은 환경에서 성장하였더라도, 쌍둥이 간 생활습관이 다르다면 질병도 상이하게 발현된다(배다정 & 박춘식, 2013).

이처럼 다양한 후성유전층은 유전체 영역이 DNA 결합 단백질에 접근 가능한지(어떤 염색체 영역이 '활성' 이고 어떤 염색체 영역이 '비활성' 인지를 결정), 또는 전사체가 단백질로 전환되었는지를 조절하는 복잡한 상호작용에 관여한다.

2.1.2. 후성유전의 발생기전

후성유전학적 발생기전에는 염색질 마크(DNA 염기 변형 및 히스톤 변형), 비코딩 RNA(non-coding RNA), RNA 염기 변형 및 고차 구조, 염색체의 핵 위치 및 전사인자(Transcription Factor, TF) 결합 패턴이 포함된다. 후성유전에서는 유전자 돌연변이와 같은 '유전자 구조' 에 대한 관찰에서 벗어나 유전자 구조의 '외부 조건' 들을 살피게 된다. 유전자 구조의 외부 조건이란, 유전자 발현을 조절하는 발생기전인 DNA 메틸화(methylation), 히스톤 변형(histone modification), 크로마틴 리모델링, Non-coding RNA(nc RNA)을 말한다(오정환 등, 2008; Bjornsson et al., 2008; 강길전, 2010; Jenuwein, 2006; Bird, 2002).

후성유전학적 조절의 개략적인 사항은 다음과 같다(그림 2).

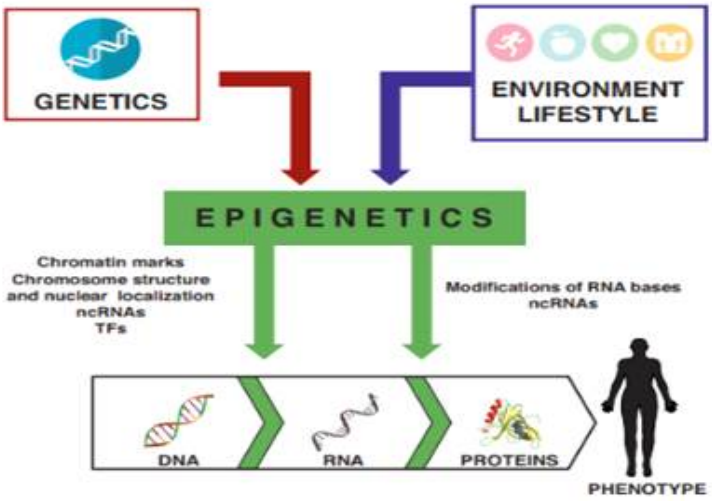

그림 2. 후성유전학적 조절
(출처: De Riso & Cocozza., 2021)

2.1.2.1. DNA 메틸화

DNA 메틸전이효소(DNA methyltransferase, DNMT)에 의해 사이토신구아닌이핵산(cytosine-guanine dinucleotide-rich, CpG)의 5번 탄소고리에 메틸그룹이 공유결합되는 유전성 후성유전학적 표지이다(Jin et al., 2011). 크로마틴에 대한 전사인자, 히스톤의 접근성을 조절함으로써 유전자 전사의 억제와 관련된다. Kumar et al(2014)에 따르면 생활습관 및 환경에 따라 발생하는 비정상 DNA 메틸화 현상은 유전자 발현에 영향을 끼치며 질병의 발생과 연관된다. 메틸화되면 유전자는 발현되지 않고, 탈메틸화되면 유전자는 발현된다(배다정 & 박춘식, 2013).

2.1.2.2. 히스톤 변형과 크로마틴 리모델링

뉴클레오솜은 크로마틴의 기본 단위로서 히스톤과 DNA의 복합체이다. 히스톤 변형은 아미노산 꼬리의 아세틸화(acetylation), 메틸화(methylation), 인산화(phosphorylation) 등을 통해 구조를 변화시키고, 유전자 전사과정을 조절한다(Kouzarides, 2007). 아세틸화는 화학작용이 일어나기 쉬운 구조로 변화되고, 환경 자극에 반응하여 더 빠르게 유전자 발현을 조절할 수 있다. 크로마틴은 저메틸화 DNA에 의하여 구조가 해체되고 히스톤과 결합한다. 이를 통해 히스톤 메틸화를 억제, 유전자 발현을 유도한다(Guan et al., 2002).

2.1.2.3. Non-coding RNA(ncRNA)

마이크로 RNA는 단일염기가닥의 non-coding RNA이다. 인체유전자의 약 30% 정도가 마이크로 RNA에 의해 조절을 받는 것으로 알려져 있다(Chen & Rajewsky, 2007). 여러 질병에서 마이크로 RNA에 의한 유전자 이상발현이 보고되어, 질병 발생 및 치료에 중요한 물질로 주목받고 있다(Sayed & Abdellatif, 2011; 송은모 등, 2020).

2.1.3. 영향요인 및 관련 문제

DNA 메틸화, 히스톤의 변형과 같은 유전자 조절기전에는 노화, 식이, 운동상태, 알코올, 흡연, 약물, 세균 및 바이러스 감염 등이 영향을 미친다. 이러한 요인은 후성유전인자의 변형을 유발한다(강길전, 2010; Bjornsson et al., 2008; 이주하 등, 2013)(그림 3).

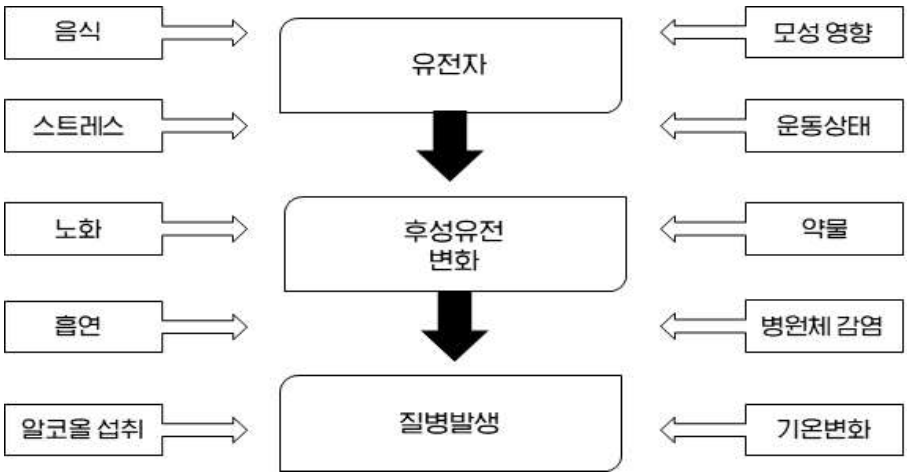

그림 3. 후성유전 변화를 유발하는 환경적 요인

2.1.3.1. 영향요인: 음식

Lawless et al(2009)에 따르면 식단에 따라 후성유전학적 변화가 유발될 수 있다. 불포화 지방산이 풍부한 식사는 유전자 변이를 일으킬 수 있는 자유 라디칼(free radical)과 산화 스트레스를 발생시킨다는 것이다. Chen & Xu(2010)은 인간 대상 연구에서 브로콜리 싹을 섭취한 후 3~6시간 이내에 말초혈액 단핵구에서 히스톤 활성 억제, 히스톤 변형이 유도됨을 확인하였다. 과일과 채소가 풍부한 식사를 할 경우 후성유전적 변화가 유발되어 항암 효과를 가져온다는 사실을 확인할 수 있었다.

2.1.3.2. 영향요인: 운동상태

Zhang et al(2011)에 따르면, 신체활동은 말초혈액 림프구의 Long interspersed nuclear element-1(이하 'LINE-1' ; 반복되는 DNA 서열로 인간 유전체의 약 17%를 구성)의 메틸화와 관련된다. 특히 말초혈액 림프구에서 LINE-1

메틸화가 높은 고령자가 허혈성 심장 질환·뇌졸중의 발병률과 사망률이 낮았다. McGee et al(2009)의 연구에서도 운동을 통해 염증 경로에 관여하는 38개 miRNA를 포함한 miRNA의 변형을 유발할 수 있다고 언급하였다.

2.1.3.3. 영향요인: 흡연

흡연으로 인한 폐암 환자의 말초혈액 림프구에서 p53 유전자의 저메틸화가 보고되었다(Woodson et al., 2001). p53은 암 억제 단백질로 다세포 생물의 세포 주기에서 암 억제자로서 암을 예방하는 데 중요한 단백질이다. p53 유전자 돌연변이에 의해 p53 단백질이 정상적으로 작동되지 못한다면 손상을 입은 DNA를 세포가 세포분열하고, 손상된 세포가 계속 분열을 할수록 손상된 DNA를 감지할 능력이 없으므로 암을 유발하는 돌연변이를 낳게 된다.

2.1.3.4. 영향요인: 환경오염

대기 중의 미립자에 대한 노출은 심혈관 질환으로 인한 발병률·사망률 및 폐암 위험 증가와 관련 있다는 것이 보고되었다. 특히 최근의 연구에서 PM10에 대한 장기간 노출이 Alu 및 LINE-1의 메틸화와 음의 관련이 있다는 사실이 확인되었다(Castro et al., 2003). 혈액 LINE-1의 메틸화 감소 현상은 암 환자 및 심혈관 질환 환자에서 발견된 바 있어, 대기오염 물질은 메틸화 변형과 질병을 유발하는 요인으로 확인되었다.

2.1.3.5. 관련 문제

이렇듯 여러 환경적 요인에 의해 후성유전 변화가 유발되고, 변화된 후성유전은 각종 질병으로 이어진다. 예를 들어 천식은 태아에서부터 성인까지 환경 노출과 유전적 요소로부터 복합적인 영향을 받는 질환이다. 임신 중 흡연을 하거나 항생

제를 사용하는 경우, 비타민 E와 아연 함유가 적은 음식을 섭취하는 것과 같은 생활습관은 소아 천식의 위험도를 증가시킨다(Tang & Ho, 2007; Devereux et al., 2006; JĘdrychowski et al., 2006).

또 다른 예로 일란성 쌍생아에서의 류마티스 질환의 차별적 발생을 살펴볼 수 있다. 일란성 쌍생아는 동일한 DNA 염기서열을 가지고 있지만, 다른 환경요인에 의해 유전자 발현 양상이 달라지게 되며 일란성 쌍생아 사이에서 질환의 발생률이 달라지는 원인이 된다. 실제로 일란성 쌍생이에서 류마티스 질환의 동시 발병 확률은 약 25%를 넘지 못한다(Deafen et al., 1992; Silman et al., 1993).

2.1.4. 후성유전 분석 관련 주요 기술

앞서 살펴본 바와 같이 후성유전 발생기전으로는 DNA 메틸화, 히스톤 변형 및 크로마틴 리모델링, non-coding RNA 등이 있다. 후성유전학 분석을 위한 기술은 꾸준히 개발되고 있다. 관련하여 여러 가지 기술 중 대표적으로 염색질 면역침전 분석(Chromatin Immunoprecipitation, ChIP), Bisulfite Sequencing(BS-Seq), 고처리분석, 염색체 교차결합기술(Chromosome Conformation Capture Techniques, 3C-based methods)을 살펴보고자 한다.

2.1.4.1. 염색질 면역침전분석

염색질 면역침전분석(Chromatin Immunoprecipitation, ChIP)은 특정 DNA/RNA 결합 단백질을 포함하는 염색질 복합체(Chromatine Complex)를 분리하거나 특정 화학적으로 변형된 염기를 가진 DNA/RNA 조각을 분리하는 기술이다. 세포 내에서 단백질과 DNA 간 상호작용을 확인하고, 특정 단백질이 특정 유전자 영역과 관련

이 있는지를 결정하는 데 활용한다. 후성유전 발생기전인 '히스톤 변형' 과 DNA 간 결합 상태를 분석하는 것을 목적으로 한다(Hamamoto et al., 2019). 단일 세포 내에서 단백질이 DNA에 교차연결되고, 단백질-DNA 복합체가 단백질에 대한 항체를 사용하여 면역침출에 의해 분리된다(Gilmour & Lis, 1985).

2.1.4.2. Bisulfite Sequencing(BS-Seq)

DNA의 메틸화 패턴을 파악하기 위하여 시퀀싱 전 단계에서 아황산 수소 소듐(Bisulfite Sequencing) 처리를 수행하는 기법이다(Fraga & Esteller, 2002). 일반적으로 BS-Seq의 목표는 사이토신(Cytocine)의 5번 탄소에 결합한 메틸화 존재 여부를 파악하기 위한 것이다.

2.1.4.3. 고처리분석

마이크로어레이 또는 차세대 시퀀싱(next-generation sequencing, NGS)와 같이 처리량이 높은 기술을 특정 후성유전학적 표지를 보유한 유전체 또는 전사체 영역과 분리하는 기술을 의미한다.

2.1.4.4. 염색체 교차결합기술

핵 DNA의 3차원 구조를 조사하기 위한 기술이다. 핵 DNA의 비인접 유전체 위치 간 상호작용을 잠금으로 고정시킨 다음, 염색질 조각화 및 DNA 연결을 수행한다. 이를 통해 3차원 공간에서 서로 가까이 있는 염색체 위치가 하나의 DNA 조각에 포착되도록 한다(Helm & Motorin, 2017).

이와 같은 후성유전 분석 기술들을 통해 다양한 후성유전학적 표지를 생성하고, 공개 데이터 세트를 검색·활용할 수 있게 되었다(Grossman et al., 2016).

2.2. 인공지능의 개념

2.2.1. 인공지능의 정의

인공지능(Artificial Intelligence, AI)이라는 용어는 1950년대 McCarthy에 의해 생성되었다. 인공지능은 알고리즘이 학습, 추론 및 문제해결과 같은 인간의 인지기능을 모방하도록 개발된 컴퓨터 과학의 한 분야를 의미한다(McCarthy et al., 2006). Kaplan & Haenlein(2019)은 인공지능을 '시스템이 외부 데이터를 올바르게 해석하고 해당 데이터에서 배우며 이러한 학습을 유연한 적응을 통해 구체적인 목표와 작업을 달성하는 능력' 으로 설명하였다. Tsang et al(2020)은 인공지능을 '컴퓨터 공학의 한 분야로 컴퓨터가 인간의 학습능력, 추론능력, 지각능력 등의 지적 능력을 모방하게 하는 것' 이라고 정의하였다.

2.2.2. 인공지능 기술영역

세계보건기구(World Health Organization, WHO)는 '인공지능 윤리와 거버넌스 지침서(Ethics and Governance of Artificial Intelligence for Health)' 에서, 인공지능을 인간이 정의한 목표들의 주어진 집합에 대해 실제 또는 가상환경에 영향을 미치는 예측, 권고, 결정을 내릴 수 있는 기계 기반 시스템으로 정의하였다(Sharma et al., 2022). '인간이 정의한 목표' 는 사람의 개입 없이 기계가 부분적으로나마 학습할 수 있도록 함으로써 달성할 수 있다(Rauschert et al., 2020). 이를 가능하게 하는 인공지능 기술의 핵심은 인공지능 모델이며, 시스템 외부 환경의 전부 또는 일부를 활용하여 환경의 구조 또는 역학을 설명한다. 모델은 전문지식 또는 데이터를 기반으로, 사람에 의하거나 자동화된 도구에 의해 작동될 수 있다(Etrel, 2019).

인공지능은 지각, 추론, 학습, 환경과의 상호작용, 문제해결, 창의력 발휘와 같은 인간의 마음과 관련된 기능을 수행하는 기계의 능력이다(Mckinsey & Company, 2023). 인공지능의 기술에 따른 영역으로는 인지(Recognition), 학습(Analytics), 추론(Inference), 행동(Performance)과 같이 표현할 수 있다. 이는 인공지능이 일련의 작업을 수행하기 위한 단계라고도 할 수 있다(그림 4).

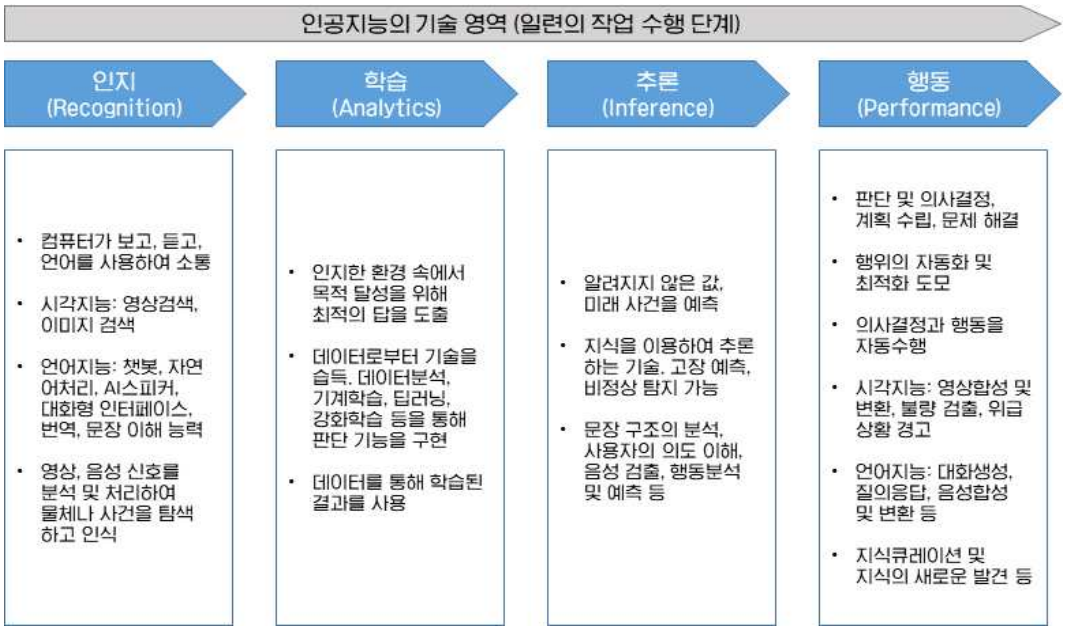

그림 4. 인공지능의 기술영역

(출처: 박소연, 이창엽, & 안찬형., 2020, 재구성; 김지선, 2019)

'인지' 는 인간 혹은 대상 시스템으로부터 받은 데이터를 컴퓨터가 이해할 수 있는 방식으로 변환하는 영역이다. 이미지, 영상데이터를 인식하여 판단하거나, 데이터를 가공하여 새로운 이미지나 영상을 생성한다. 시각 지능의 학습기술로는 입력된 이미지 데이터에서 객체를 인식하는 것에 있다. 객체 영역을 식별한 후

분리된 객체 영역의 특징을 분석 및 인식한다. 이 과정에서 다량의 이미지 데이터 학습이 필요하며, 이때 지도 또는 비지도 학습의 방식이 활용된다(문상선, 2023). 언어지능 기술은 인간이 사용하는 일상적인 방법으로 언어를 이해하고 대화하는 것으로, 자연어처리(Natural Language Processing, 이하 'NLP')를 중심으로 연구가 진행되고 있다(elastic, n.d.). 이외에도 여러 종류의 센서를 통해 기계가 이해할 수 있는 디지털 신호로 변환하여 탐색하고 인식하도록 한다.

'학습'은 보유한 데이터를 기반으로 판단을 위한 패턴을 학습하는 영역이다. 데이터 학습 목적과 형태에 따라 지도학습, 비지도학습 등의 기법을 활용한다. 이를 통해 복잡하고 비구조화된 데이터를 처리하고 패턴을 식별하며, 정확한 예측을 수행함으로써 학습 및 분석 능력을 강화할 수 있다(Qlik, n.d.).

'추론'은 학습된 패턴을 바탕으로 상호작용을 통해 입력된 데이터의 의미를 판단하거나 결과를 예측하는 것이다. 예를 들어 기계학습의 교육 단계에서 개발자는 분석할 데이터 유형에 대해 필요한 모든 것을 '학습'할 수 있도록 모델에 큐레이션(데이터를 분석하는 데 있어서 효율적으로 데이터를 검색 및 활용, 판단하는 작업)된 데이터 세트를 공급한다. 그런 다음 추론 단계에서 모델은 실행 가능한 결과를 얻기 위해 실시간 데이터를 기반으로 '예측'을 수행할 수 있게 되는 것이다(xilinx, n.d.).

마지막으로 '행동'은 추론 결과를 바탕으로 물리적인 행동을 수행하거나 시스템의 처리를 발생시켜 실제적인 작업을 수행하도록 하는 것이다. 상황 판단에 근거하여 실제로 영상 및 이미지를 조작하는 기술이 주를 이룬다. 스마트폰 카메라 이미지 보정 기능, 딥페이크 등을 이러한 예로 들 수 있다. 학습 및 추론 단계에서 이해한 자연어에 대응하여 응답을 생성하며, 준비된 예제를 기반으로 대응하는 방식을 주로 사용한다(박소연, 이창엽, & 안찬형, 2020).

2.2.3. 인공지능의 주요 기법

Jung & Park(2022)에 따르면, 인공지능은 인공지능, 기계학습(Machine Learning, ML), 딥러닝(Deep Learning, DL)과 같이 구분할 수 있다. 이외에도 인공지능에는 다양한 기법들이 존재한다(그림 5).

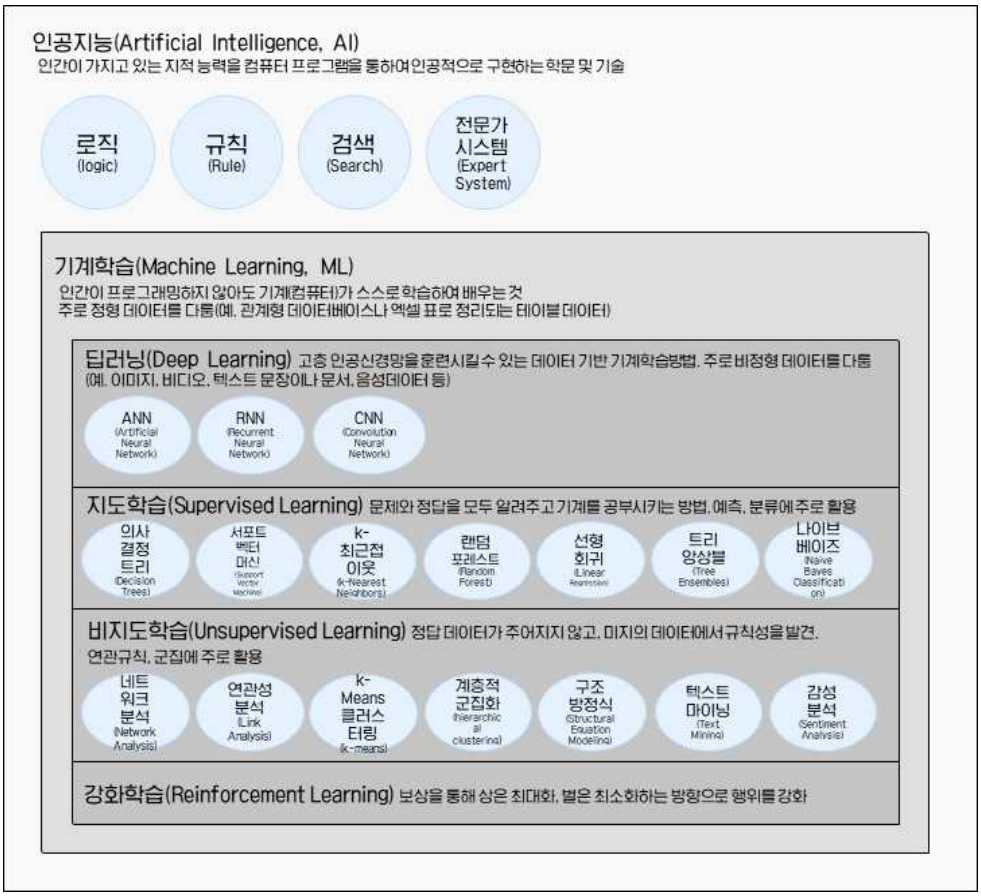

그림 5. 인공지능 구분에 따른 세부 기법

(출처: 박소연, 이창엽, & 안찬형., 2020; 조민호, 2021)

유전학 및 후성유전학에서 주로 활용하는 인공지능 기법은 '기계학습(Machine Learning)'이다(Rajkomar et al., 2019). '기계학습'은 컴퓨터가 데이터에서 패턴을 자율적으로 찾아내어 특정 도메인에 대한 지식을 개발할 수 있도록 하는 인공지능 분야를 말한다(Samuel, 1959). 컴퓨터는 경험을 통해 작업의 성능을 향상함으로써 학습하게 된다(Mitchell, 1997). 일반적으로 가설 주도적 추론보다 예측 정확도에 중점을 두며, 일반적으로 많은 고차원 데이터 세트에 초점을 맞춘다. 특히 빅데이터에 관한 관심 증대에 따라 기계학습 기술 기반의 알고리즘은 일반적으로 사용되는 통계적 접근 방식으로 해결하기 어려운 문제에 더 많은 도움이 될 수 있다(Rasmussen & Williams, 2006). 기계학습은 데이터에서 지식을 추출하기 위해 몇 가지 가정을 세우고, 규칙을 추출하여, 도메인을 모델링 한다. 실제 데이터와의 비교를 통해 정교화 단계를 거치고, 생성된 최종 모델은 시스템 자체에서 미지의 데이터를 분류하거나 연구자가 새로운 지식을 습득하고 분석된 도메인에 대한 통찰력을 도출하는 두 가지 방식으로 사용될 수 있다(Ghahramani, 2015; Perakakis et al., 2018).

기계학습은 학습 패러다임에 따라 '지도학습(Supervised Learning) ', '비지도학습(Unsupervised Learning)', '강화학습(Reinforcement Learning)'으로 구분된다. 지도학습은 주로 예측, 추정, 분류 작업에 적용되며, 비지도학습은 주로 패턴 및 구조의 발견, 그룹화, 차원의 축소, 클러스터링에 적용된다. 이외에도 기계학습의 하위 유형으로 '딥러닝(Deep Learning)' 기술이 있다.

2.2.3.1. 지도학습(Supervised Learning)

지도학습 알고리즘은 알려진 결과를 가진 데이터에서 학습하여 알려진 결과가 없는 데이터에 대한 예측을 수행한다. 학습 단계에서 알고리즘은 제공된 예시로 모델을 개발하고, 생성된 모델을 알려진 결과가 없는 데이터에 적용하여 평가한

다. 예측과 실제 결과 간 일치가 높을수록 모델의 성능이 높아진다(Arslan et al., 2021). 지도학습 알고리즘은 분류 및 예측 작업에 효과적으로 활용된다. 흔히 사용되는 알고리즘에는 선형 또는 로지스틱 회귀(linear or logistic regression), 서포트 벡터 머신(Support Vector Machine, SVM), 랜덤 포레스트(Random Forest, RF), 최소 절대 수축 및 선택 연산자 회귀(least absolute shrinkage and selection operator regression, LASSO) 등이 있다. 지도학습은 질병이 있는 개인과 건강한 개인을 분류하는 강력한 방법을 제공하지만, 다음과 같은 한계를 가진다. 첫째, 모델을 개발하기 위한 클러스터를 정의하기 위해 사용자 입력을 필요로 한다. 둘째, 데이터의 품질에 민감하므로 올바른 예측 결과를 불러오기 위해, 올바른 레이블 지정이 필요하다. 셋째, '과적합'에 민감하다. 즉, 훈련 데이터에서는 잘 작동되지만 다른 외부 데이터 집합에 적용되었을 때 성능이 저하될 수 있다(Rauschert, 2020).

2.2.3.2. 비지도학습(Unsupervised Learning)

비지도 학습은 작업을 수행하기 위해 별도의 레이블을 설정하는 것이 요구되지 않는다. 유사성에 따라 데이터 객체를 분류한다. 주로 데이터 집합을 '클러스터'라는 그룹으로 분할하여, 동일한 클러스터 내 관측치와 다른 클러스터 내 관측치 간 유사성을 최대화하는 것을 목표로 한다. 데이터 집합 내 변수 간의 상관성을 확인하지만, 알고리즘에 의해 식별된 관련 데이터 집합의 타당성과 중요성을 할당하는 능력은 없으므로 의미와 라벨을 부여하기 위해 인간의 개입(수동 검사)이 요구된다(Rauschert, 2020). 그러나 숨겨진 특성을 가진 클러스터를 식별할 수 있다는 점이 가장 큰 장점으로 작용한다. 데이터의 내재 그룹에 따라 클러스터링하며, 비지도 학습에서 사용되는 일반적인 방법에는 k-평균 클러스터링(k-means clustering) 및 계층적 클러스터링(hierarchical clustering), 주성분 분석(principle component analysis) 및 부분 최소 제곱 판별 분석(partial

least squares discriminant analysis)을 포함한다(Tacra et al., 2007). 비지도 학습 알고리즘은 특히 많은 양의 데이터를 보유한 데이터 집합에서 패턴을 감지하는 데 유용하여, 임상 연구대상에서 유사한 특성을 가진 환자 그룹을 찾을 때 비지도 학습은 효과적으로 작용할 수 있다(Bae et al., 2014). 그러나 비지도 학습 알고리즘은 잡음에 민감하여 데이터 집합 내 무관한 데이터가 많은 경우 클러스터링에는 오류가 발생할 수 있다(Krittanawong et al., 2017).

2.2.3.3. 딥러닝(Deep Learning)

딥러닝은 기계학습(Machine Learning, ML) 기법을 바탕으로, 생물의 신경계를 모방한 인공 신경망의 하나로써 신경망의 학습 수준을 높이는 모델이다. 주어진 데이터로부터 패턴이나 특성을 학습하여 새로운 데이터에 대해 작업을 수행해 낼 수 있도록 하는 알고리즘을 의미한다(이한상 등, 2014; Esteva et al., 2019). 인간에 의해 명시적으로 지정되는 것이 아니라 대규모 집합에서 자동으로 파생되는 통계적 데이터 규칙을 사용하여 알고리즘의 입력을 출력으로 변환한다. 인공지능의 최근 부흥을 주도한 기술이며 기존의 기계학습 기술을 능가한다(Yu et al., 2018). 지도학습과 비지도학습의 두 가지 방식으로 학습기술이 구분되어 있다. 그러나 학습과정에서 특징을 자체적으로 추출하고 학습을 수행함으로써 대상의 특성과 관계없이 일반적인 모델링이 가능하다(이한상 등, 2014). 이에 의료영상 판독 등 이미지나 영상데이터를 처리하는 분야에서 가장 크게 활용되고 있으며, 이미지 데이터를 입력하여 이미지 특정 패턴을 찾아내거나 비슷한 특성끼리 묶는 등의 작업을 수행할 수 있다. 딥러닝은 무한한 데이터를 모두 기억할 수 있으며, 제한적 공간에서의 강화학습·잠재인자(latent factor) 모델링·데이터 정리, 처리 및 정보 생산 등의 성과를 낼 수 있다.

제3장 후성유전 분야에서의 인공지능 적용현황

3.1. 개요

후성유전학(epigenetics)적 변화는 메틸화, 히스톤 변형, miRNA 등 신호를 이용하여 DNA 염색질의 구조적 변화에 따라 유전자가 조절되는 현상을 의미한다. DNA의 저메틸화는 발암 유전자를 활성화하는 반면 DNA의 과메틸화는 종양 억제유전자를 불활성화하며, 고위험 질환군 식별을 위한 생체지표로 사용될 수 있다(Lee et al., 2022). 이러한 특성을 가진 후성유전학적 변화는 다양한 질병 상태와 관련된 주요요인이며, 향후 진단 및 치료의 민감도와 특이도를 높이는 기회를 제공한다. 생활습관, 환경과 유전정보 간 연결고리는 건강 요소의 90%를 차지하기에 생활습관과 환경으로 인해 발현된 유전자 변이가 다음 세대에게 유전되는 후성유전 정보를 고려하는 것은 매우 중요하다. 그러나 후성유전을 분석하여 얻은 데이터는 Omics 데이터(생물학적 데이터)의 전형적 구조를 가진다(Xu & Jackson, 2019).

최근에는 인구 수준의 시퀀스 데이터를 풀링하고 유전체 데이터를 형질 정보, 임상 기록 및 기타 다양한 Omics 데이터와 연결하는 노력이 커지고 있다. 이러한 데이터는 생리적 측정치, 의료 이미지 처리 데이터(CT, MRI 스캔), 기타 임상 정보를 포함할 수 있다. Omics 및 임상 데이터의 통합 분석은 개인 맞춤형 의학을 위한 새로운 생물 의학적 발견을 촉진하는 데 중요하며, 연구자 및 임상 서비스에 복잡한 분석과 계산을 요구한다(Sobia Raza, 2020). 이에 대규모의 이질적이며 고차원이라는 특성을 가진 데이터 분석을 다룰 수 있는 컴퓨팅 접근 방법에 대한 수요가 계속 증가하고 있다(Sobia Raza, 2020).

이런 측면에서 후성유전적 정보를 활용하여 아직 실험이 수행되지 않은 사례의 결과를 예측할 수 있는 인공지능 기반 접근 방법을 활용하는 것은 매우 중요하다(Aroa & Tollefsbol, 2021). 인공지능 기술은 데이터에 명시적 규칙을 지정하지 않고도 새로운 발견을 쉽게 할 수 있다(De Riso & Cocozza, 2021). 매우 큰 데이터 세트 및 여러 데이터 유형을 입력으로 처리할 수 있는 능력은 헬스케어 및 유전의학의 발전을 도울 수 있으며, 대규모 건강 데이터를 분석하는 데 도움을 줄 수 있다. 시간적·재정적 비용 소모를 감축하고 의료서비스를 활성화할 수 있다. 이를 통해 환자 만족과 치료의 질 향상, 효과적인 정밀의료 개입을 가능하게 할 수 있다(Rauschert et al., 2020; Habuza et al., 2021).

이에 본 연구에서는 후성유전 분야 연구에 적용된 인공지능 기술을 살펴보고자 하였다. 국내 데이터베이스의 경우 0개, 국외 데이터베이스의 경우 71개 문헌을 추출하였으며, 수기 검색을 통해 6개 문헌을 추가로 검토하고자 하였다. 참고문헌 관리를 위해 EndNote 21을 사용하였다. 제목과 초록을 바탕으로 적격성을 결정하였으며 초록만으로 적격성을 확인할 수 없는 경우 전문(Full-Text)의 내용을 확인하였다. 77개 문헌 중 선정기준에 부합하지 않는 문헌 51개를 제외하여 최종 24개의 문헌을 분석하였다.

최종 선정된 문헌은 엑셀 스프레드시트를 사용하여 다음과 같이 정보를 추출하였다: 1) 일반정보(저자, 발행연도, 국가, 연구목표, 연구유형 및 설계), 2) 인공지능의 유형 및 활용(사용된 인공지능 기술 및 알고리즘, 작업목표), 3) 실행과정(연구초점, 관련 질병, 후성유전 발생기전). 문헌 선정의 신뢰도를 높이기 위해 본 연구자 외에도 주심 교수 1인, 부심 교수 2인을 팀으로 구성하여 추출 결과에 대해 논의하였다(그림 6).

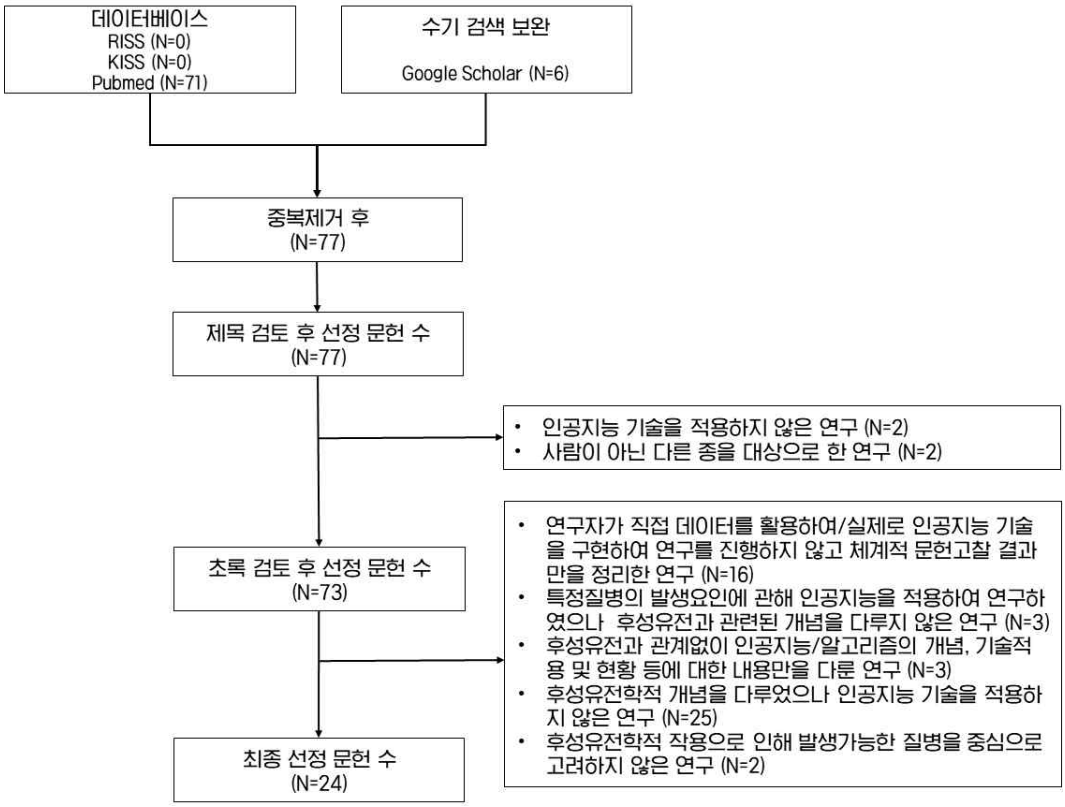

그림 6. 문헌의 선정 도식도

3.2. 인공지능 적용현황 분석

앞서 설정한 첫 번째 연구문제에 의거 주제범위 문헌고찰에 포함된 24편의 문헌에 대한 분석 결과를 다음과 같이 정리하였다.

3.2.1. 후성유전 분야에서 실제 인공지능 기술을 구현한 연구의 특징

표 2. 주제범위 문헌고찰에 포함된 문헌 개요

번호	저자	출판연도	국가	연구목표	연구설계
1	Imgenberg-Kreuz et al	2019	스웨덴	전신홍반루푸스, 일차쇼그렌증후군 관련 DNA 메틸화 유발 유전자 확인	임상시험
2	J Orozco et al	2018	미 국	뇌종양의 최적 치료를 목적으로 뇌종양 유형에 따른 후성유전 특징 확인 및 뇌전이 분류 모델 구축	코호트연구
3	ElHefnawi et al	2013	이집트	간암에서 miRNA의 역할을 재조명 하기 위한 표적 예측	메타분석
4	Min et al	2018	싱가포르	어린이 대상 수족구병 타액 샘플의 miRNA 프로파일 확인 및 miRNA 기반 진단모델 개발	임상시험
5	Yang et al	2018	중 국	폐선암 표본에서 흡연과 관련된 특정 유전자를 식별하고 메커니즘 탐구	메타분석
6	Chakravarthy et al	2016	영 국	인유두종바이러스(HPV)가 편평세포암을 유발하는지 확인하고 생존 이익 확인	메타분석
7	Huan T et al	2022	미 국	DNA 메틸화를 임상적 위험인자와 통합하여 예측 모델 구축, 사망률 예측 향상	메타분석
8	Cheng et al	2023	중 국	전립선암 예후 예측을 위한 long non-coding RNA 식별	임상시험
9	de Gonzalo-Calvo et al	2020	독 일	예측모델이 혈액투석 환자의 의료적 의사결정에 유용한 정보를 제공하는지 확인	임상시험

표 2. 주제범위 문헌고찰에 포함된 문헌 개요(계속)

번호	저자	출판연도	국가	연구목표	연구설계
10	Bahado-Singh et al	2022a	미 국	태반 조직의 후성유전학적 차이와 자폐증 발달과의 연관성 확인	환자-대조군 연구
11	Tran et al	2022	미 국	중추신경계 종양 분류에서 지도학습모델의 정확도 향상을 위해 메틸화 데이터에 대한 기계학습 적용을 탐구	코호트연구
12	Bahado-Singh et al	2022b	미 국	대동맥협착에의 후성유전학적 변화 확인, 질병발생의 잠재적 예후 확인	환자-대조군 연구
13	Yu Y et al	2021	중 국	기계 학습을 활용하여 초기 단계인 침윤성 유방암 환자에서 액와림프절 상태를 평가하는 접근법 개발	임상시험
14	Diboun et al	2021	카타르	뼈의 파젯병에서 DNA 메틸화 프로파일의 파젯병 관련 변형에 의한 역할을 확인	메타분석
15	Karisola et al	2021	핀란드	달걀을 이용한 경구 면역요법을 활용하여 탈감작을 유도하는 세포 매개 분자 메커니즘을 확인	환자-대조군 연구
16	Aref-Eshghi et al	2018	캐나다	표준 임상 진단을 보완하기 위해 증후군 특이적 바이오마커를 제공할 수 있는 DNA 메틸화 후성유전 정보를 확인	임상시험
17	Shokhirev & Johnson	2022	미 국	데이터세트에 기계학습모델을 적용하여 알츠하이머를 유발하는 생물학적 과정 이해, 알츠하이머 진단을 위한 분자유형 입증	메타분석
18	Hess et al	2020	미 국	양극성장애와 조현병 관련 유전학적 변화 확인, 바이오마커와 병원성 기전을 확인	메타분석
19	Kalyakulina et al	2022	러시아	DNA 메틸화 데이터 기반의 환자 분류 워크플로우 제안	메타분석
20	Bendifallah et al	2022	프랑스	자궁내막증 환자 구분, 혈액기반 miRNA 진단 시그니처 개발	임상시험

표 2. 주제범위 문헌고찰에 포함된 문헌 개요(계속)

번호	저자	출판연도	국가	연구목표	연구설계
21	Arabyarmohammadi et al	2022	미 국	골수아세포의 염색질 패턴에서 재발 및 생존 예측	임상시험
22	Bahado-Singh et al	2023	미 국	태아 선천성 심장결함의 최소 침습적 검출을 위해 인공지능 분석 사용	코호트연구
23	Lin et al	2020	미 국	알코올 의존성에 대한 날트렉손(naltrexone, NTX) 치료가 DNA 메틸화에 미치는 영향 확인	임상시험
24	Huang et al	2021	싱가포르	ART 보조 생식법이 아동건강결과 및 심혈관 대사 결과에 미치는 영향을 추정	임상시험

검토된 문헌의 42%(10편/24편)가 2022년과 2023년에 발표되었다(그림 7). 미국에서 진행된 연구가 10편으로 가장 많았으며, 중국 3편, 싱가포르 2편 순으로 많았다(그림 8). 연구설계는 임상시험 10편으로 가장 많은 비중을 차지하였으며, 메타분석(8편), 코호트(3편), 환자-대조군 연구(3편) 순으로 많았다(그림 9).

그림 7. 검토 문헌의 연도별 편 수

그림 8. 검토 문헌의 국가별 편 수

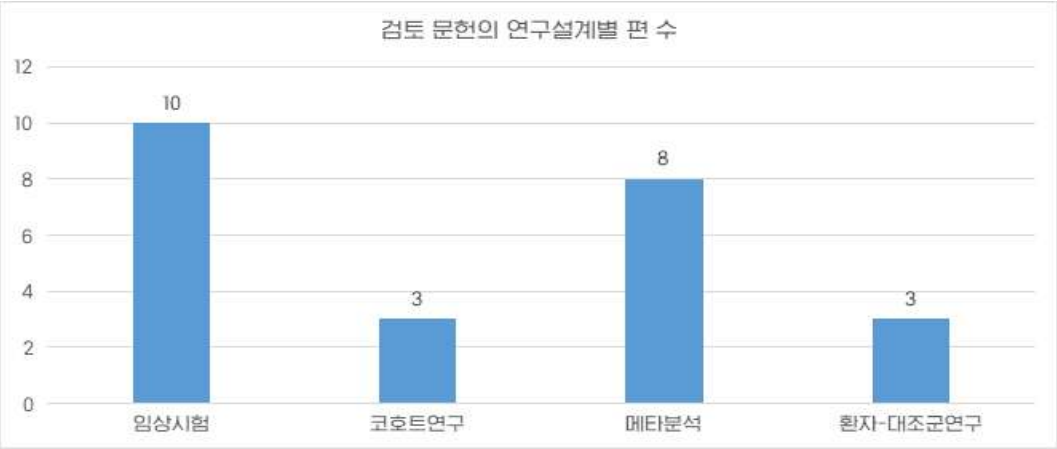

그림 9. 검토 문헌의 연구설계별 편 수

3.2.2. 연구에서 구현된 인공지능 기술과 활용 데이터베이스 및 데이터 세트

앞서 설정한 두 번째 연구문제에 의거, 주제범위 문헌고찰에 포함된 연구에서 구현된 인공지능 기술, 활용 데이터베이스 및 데이터 세트를 다음과 같이 정리하였다.

3.2.2.1. 구현된 인공지능 기술

주제범위 문헌고찰에 포함된 모든 연구에서 구현된 인공지능 기술은 기계학습(Machine Learning, ML)이었다. 구현된 인공지능 알고리즘은 랜덤 포레스트(Random Forest, RF)가 가장 많았으며(12편/24편, 50%), 서포트 벡터 머신(Support Vector Machine, SVM)이 두 번째로 많았다(9편/24편, 38%). 랜덤 포레스트 알고리즘만을 활용한 연구는 3편이었으며, 서포트 벡터 머신만을 활용한 연구는 4편이었다. 단독 알고리즘을 활용한 7편을 제외한 17편의 연구에서는 랜덤 포레스트 또는 서포트 벡터 머신과 함께 다른 기계학습 알고리즘을 활용하고 있었다.

'랜덤 포레스트' 알고리즘의 변형인 '랜덤 생존 포레스트(Random survival forest, RSF)'이 질병으로 인한 예후 확인(생존분석)을 위해 활용되었다. 이외에도 '서포트 벡터 머신'의 변형 알고리즘인 '커널 서포트 벡터 머신(Linear-kernel support vector machines, SVM)', '방사형 커널 서포트 벡터머신(radial-kernel, SVM)', 'Classification and Regression Tree(CART)' 등이 활용되고 있었다. 특히 암과 관련된 연구에서는 'CIBERSORT' 알고리즘을 활용하고 있었다. 기타 알고리즘의 정의와 특성은 <부록 1>과 같다.

3.2.2.2. 활용된 데이터베이스 및 데이터 세트

주제범위 문헌고찰에 포함된 연구에서 활용한 데이터 세트는 각 병원에서 수집한 환자 데이터 또는 연구자가 별도로 수집한 환자 데이터를 활용한 경우가 대부분이었다. 수집한 데이터는 github(github.com/jhuang35/ivf_growth/), Zenodo (10.5281/zenodo)를 통해 게시되기도 하였다. 그러나 대부분의 연구들에서는 NCBI의 Gene Expression Omnibus (이하 GEO)를 통해 게시된 여러 자료(특히, 종양 참조 코호트 데이터)를 활용하고 있었다. 암과 관련된 연구에서는 GEO 뿐 아니라 The Cancer Genome Atlas(TCGA), TCGA Data Portal의 코호트 데이터를 함께 사용하기도 하였다. 이외에도 국가별 보건부가 수집한 데이터베이스를 활용하거나, 유전자 발현 데이터 확인을 위해 Interferome v2.01, ToppGene Suite, PhenomiR, ArrayExpress, Infinium HumanMethylation 450 BeadChip(Illumina, USA)를 활용하기도 하였다.

3.2.3. 연구에서 구현된 인공지능 기술의 적용 목적

문헌고찰에 포함된 연구들은 궁극적으로 (1) 질병 예측을 위해 인공지능을 적용한 연구, (2) 환자 분류를 위해 인공지능을 적용한 연구, (3) 치료의 차별적 예후를 확인하기 위해 인공지능을 적용한 연구와 같이 인공지능 적용 목적을 구분해볼 수 있었다.

3.2.3.1. 질병 예측을 위해 인공지능을 적용한 연구

Imgenberg-Kreuz et al(2019)의 연구에서는 전신홍반성루푸스(systematic lupus erythematosus, 이하 'SLE') 환자, 일차 쇼그렌 증후군(primary

Sjögren's syndrome, 이하 ‘pSS’) 환자 및 건강한 대조군 간의 DNA 메틸화를 교차분석하여 질병특이적 변화를 감지하는 것을 목표로 하였다. SLE와 pSS의 정확한 병인은 결정하기 어려우나, 유전적·환경적 유발요인·후성유전학적 메커니즘이 질병 발병에 기여하는 것으로 알려져 있다. 환자군 데이터는 스웨덴 Uppsala University Hospital의 Department of Transfusion Medicine에 방문하는 환자 데이터를 활용하였다. 기타 활용 데이터는 유전자 분류를 위해 Interferome v2.01 database를 활용, 기능성 유전자 세트 농축 분석은 ToppGene Suite database를 사용하였다. 유전자 전체 DNA 메틸화 패턴 중 질병을 유발하는 표적 특징을 추출하고 질병 상태를 분류하기 위해 기계학습 기술을 적용하였다. 세부 알고리즘으로는 ‘랜덤 포레스트(Random forest, RF)’를 활용하였으며, 예측 능력과 성능을 향상시키기 위해 선형회귀분석(linear regression)을 추가로 수행하였다. 연구결과 SLE와 pSS를 유발하는 후성유전학적 구조를 관찰하고 유사한 병원성 메커니즘을 확인하였다. 즉 게놈 전체 DNA 메틸화 패턴에서 SLE와 pSS 질병을 유발하는 표적 특징을 추출할 수 있었다.

J Orozco et al(2018)의 연구는 뇌종양의 최적 치료를 목적으로 정상, 일차 및 전이성 뇌종양 유형에 따른 후성유전학적 특징을 확인하고 뇌 전이를 분류하는 3단계 DNA 메틸화 기반 분류기 구축하는 것을 목표로 하였다. 종양의 치료법에는 수술, 방사선치료, 화학요법, 표적치료 및 면역요법과 같은 전신 약물치료를 포함한다. 이러한 다양한 치료를 환자 맞춤형으로 제공하기 위해서는 뇌종양에 대한 정확한 진단이 필수적이다. 이에 J Orozco et al(2018)은 전이성 뇌종양의 후성유전학적 특성의 본질적 차이를 확인하기 위해 미세절개한 뇌종양 조직에 대한 DNA 메틸화 처리를 수행하였다. 이를 통해 뇌종양 유형별로 다른 DNA 메틸화 특성을 확인하였다. 이러한 메틸화 패턴의 차이를 기반으로 ‘랜덤 포레스트(Random forest, RF)’ 알고리즘을 활용하여 뇌종양 조직을 효율적으로 식별하기 위한 분류 모델을 구성하였다. 연구결과를 통해 뇌종양 유형에 따라 환자를 분류하고,

뇌종양의 분자적 특징을 파악하여 예후를 결정하고 최적의 치료 방법을 선택하는 데 도움을 줄 수 있었다. 연구에서 활용한 환자 데이터는 Providence Saint John's Health Center(미국 샌타모니카), Melanoma Institute of Australia(호주 시드니) 및 스웨덴 의료 센터(미국 시애틀)에서 수술 가능한 원발성 또는 전이성 뇌종양 환자 165명의 자료를 수집하였다. 관련 데이터는 NCBI의 Gene Expression Omnibus(GEO)에 저장되었다.

ElHefnawi et al(2013)은 간세포암종(Hepatocellular carcinoma, 이하 'HCC') 환자의 예후와 생존율 향상을 위해 간세포암종의 조기 진단을 통한 바이오마커 및 약물표적, 치료개입 전략 수립을 연구의 필요성으로 제시하였다. 이에 HCC에서 차별적으로 발현되는 miRNA 표적 유전자 예측 및 분석하고자 하였다. 기계학습, 정렬, 통계기술을 결합한 목표 예측 도구를 활용하여 miRNA 대상 식별 및 표적 예측을 시도하였다. 데이터베이스로는 PhenomiR database (www.mips.helmholtz-muenchen.de/phenomir/)를 활용하였다. miRNA 예측을 위해 TargetScan 5.1, PicTar, DIANA-microT v3.0, miRDB 및 miRanda의 5개 프로그램을 활용하였다. 이 중 miRanda 프로그램에서 기계학습 알고리즘인 '서포트 벡터 머신(Support Vector Machine, SVM)'을 활용하였다. 연구결과 HCC와 관여된 유전자 표적 특징을 확인할 수 있었다.

Min et al(2018)에서는 타액 내 miRNA를 추출하여 수족구병을 탐지하는 모델을 개발하였다. 수족구병 환자와 건강한 대조군 사이에 표적 침상 miRome 분석을 수행하였다. 수족구병의 새로운 바이오마커를 이용한 감염병 진단 워크플로우 개발을 위해 기계학습 알고리즘인 '서포트 벡터 머신(Support Vector Machine)'을 사용하고자 하였다. 알고리즘 지도를 위한 훈련 데이러토 싱가포르 코호트 데이터를 활용하였다. 환자 데이터는 2012년 8월부터 2016년 2월까지 Kandang Kerbau 여성아동병원에서 확보한 총 35건의 수족구병 의심자 대상 인후면봉 및 타액 임

상 검체 데이터를 활용하였다. 연구결과 수족구병과 관련된 주요 표적 특징 miRNA를 도출할 수 있었다. 연구자들은 수족구병 진단을 목적으로 개발된 기계학습 모델을 통해 감염병 예방과 학교 및 인구집단 대상의 예방적 감시가 가능하다는 점에서 유용하다고 언급하였다.

Yang et al(2018)의 연구에서는 폐선암(Lung adenocarcinoma, 이하 'LAC') 표본에서 흡연과 관련된 특징 유전자를 확인하고 근본적인 메커니즘을 탐구하는 것을 목표로 하였다. 생물학적 데이터로서 유전자 발현 데이터 세트를 수집하고, 메타분석을 활용하여 발현 데이터를 처리하였다. 이를 통해 흡연자와 비흡연자의 표본에서 차별적으로 발현된 유전자(Differentially expressed genes, DEGs)를 확인하였다. 활용 데이터베이스로는 NCBI Gene Expression Omnibus(GEO) Database(www.ncbi.nlm.nih.gov/geo/)를 활용, 'lung adenocarcinoma', 'Homo sapiens', 'smoke'와 같은 키워드를 적용하였다. 발현 유전자는 DNA 메틸화 변화를 통해 염색체 불안정성에 영향을 미치고 비정상적인 유전자 발현에 영향을 미치는 유전자로 확인되었다. 발현 유전자 선택을 위한 분석, 즉 특징 유전자를 이용하여 흡연자와 비흡연자의 데이터를 분류하기 위한 유전자 식별 도구로서 기계학습 기반의 알고리즘인 '서포트 벡터 머신(support vector machine)'을 활용하였다. 연구결과 LAC 질병을 가진 흡연자들에게서 LAC을 유발하는 유전자들이 확인되었고 그 중 일부는 폐암을 유발하는 유전자로 확인되었다. 연구자들은 흡연과 연관된 LAC의 유전적 메커니즘을 뒷받침하는 연구결과라고 언급하였다.

Chakravarthy et al(2016)은 인유두종바이러스(Human Papilloma Virus, 이하 'HPV')가 구강인두편평세포암의 위험 인자라는 사실을 바탕으로, 유전자 특성에 기반을 둔 분류자를 사용하여 HPV가 구강인두편평세포암이 아닌 암종에 인과적 역할을 하는지 여부를 확인하고 이러한 종양의 예후에 HPV가 어떠한 영향을 미치는지를 예측하고자 하였다. 활용 데이터로는 유전자 발현 마이크로어레이 데

이터의 경우 ArrayExpress(www.ebi.ac.uk/arrayexpress) 에서 수집, TCGA HNSCC 분자 데이터와 임상 메타 데이터는 TCGA Data Portal(https://tcga-data.nci.nih.gov/tcga/)과 캘리포니아 대학교 산타크루즈, Cancer Browser(https://genome-cancer.ucsc.edu/)를 활용하였다. 분석 결과 HPV 표본에서 총 159개의 유전자가 차별적으로 발현되었으며, 발현된 특징 유전자가 다양한 세포 계통에서 HPV 유전자에 의해 조절되는 것이 확인되었다. 이는 HPV 관련 암에 대해 발표된 유전자 발현특성과 일치하는 결과로 나타났다. 또한, 유전자 데이터 집합이 HPV 상태를 정확하게 예측할 수 있는지 검증하기 위해 기계학습 알고리즘 중 'K-최근접 이웃(K-Nearest Neighbors)', '랜덤 포레스트(Random forest, RF; Salford Systems, San Diego, CA)'을 사용하였다. 연구결과 기계학습 모델은 구강인두편평세포암과 구강인두편평세포암이 아닌 암종에 대한 예측 분류를 가능하게 하였다. 구강인두편평세포암종이 아닌 암종의 주요 특성 유전자와 HPV의 주요 특성 유전자 간 일치하는 유전자 발현특성을 공유한다는 사실을 확인하였으며 HPV가 구강인두편평세포암이 아닌 암종과도 관련되는 유전자 발현특성을 가진다는 사실을 밝혀내었다.

Huan T et al(2019)은 후성유전학적 발생기전 중 DNA 메틸화에 초점을 맞추어, 암 또는 심혈관 질환 사망 예측을 위한 정보 바이오마커 유용성을 확인하기 위해 기계학습 모델을 적용하였다. 10년의 평균 추적관찰을 통해, 추적기간 동안 심혈관 질환(Coronary Vascular Disease, 이하 'CVD') 사망자 1235명과 암 사망자 868명을 포함한 모든 원인으로 인한 사망자 4314명을 산출하였다. 모든 원인 사망률에 대한 조상 계층화 메타분석 결과 유럽과 아프리카 혈통에서 임상적 위험요인이 높게 나타났다. 위험요인을 조정한 후 DNA 메틸화 기반 심혈관 질환 예측 모델을 개발하였다. 그 결과 DNA 메틸화와 CVD 사망률 및 암 사망률과 관련하여 메타분석 수행 결과 유의한 양의 상관관계가 있음을 확인하였다. DNA 메틸화 데이터를 사용하여 사망위험을 예측할 수 있는지 확인하기 위해 CVD 사망률 및 암

사망률에 대한 기계학습 예측모델 수준을 평가하고자 하였다. 이를 위해 'Elastic-coxph', '랜덤 생존 포레스트(Random survival forest, RSF)', 'Cox-nnet' 및 'DeepSurv' 알고리즘을 활용하였다. 연구결과 DNA 메틸화 데이터를 임상적 위험인자와 통합하여 훈련한 예측모델은 임상적 위험인자만으로 훈련한 모델에 비해 모든 원인 사망률과 CVD 사망률에 대한 예측에서 낮은 효과성을 보였고, 암 사망률에 대한 예측에서는 높은 효과성을 보였다. 연구자들은 사망률 기반 예측모델은 원인 및 특정 사망위험을 평가하고 치료 전략을 개발하는 데 유용한 임상 도구로 활용될 수 있다고 언급하였다.

Cheng et al(2023)에서는 의료개입 및 의료의 정밀성 향상을 목표로 전립선암의 예후 및 치료반응 예측을 위한 유전자 표적 특징을 확인하고자 하였다. 이에 전암 세포 사멸 과정인 cuproptosis 예후 예측 가치와 면역 미세환경, 면역 체크포인트 및 일부 흔한 호르몬 치료 약물과의 관련성을 평가하였다. 이를 검증하기 위한 생물학적 데이터로는 'The Cancer Genome Atlas(TCGA)' 코호트를 활용하였으며, 선행연구를 통해 13개의 cuprotosis 관련 유전자를 확인하고 관련하여 차별적으로 발현된 특징 유전자를 도출하였다. 전립선암의 발생을 예측하기 위한 훈련 모델 구축을 위해 기계학습 기술을 활용하였으며 관련 알고리즘으로는 '라쏘 회귀(Lasso Regression)', 'stepwise COX' 을 사용하였다. 연구결과 전립선암의 생존 예후를 정확하고 안정적으로 예측할 수 있는 기계학습 알고리즘의 구축을 통해 새로운 cuproptosis 관련 long non-coding RNA(lncRNA) 시그니처를 확인할 수 있었다. 예후징후는 여러 공통적인 임상적 특징, 면역세포 침윤, 면역관련 기능, 유전자 변이, 약물 민감도 등과 밀접한 관련이 있어 전립선암 환자의 정밀한 치료와 임상적 결과를 개선하는 데 유용하게 작용할 수 있다.

de Gonzalo-Calvo et al(2020)은 혈액투석을 하는 말기신장질환환자를 대상으로 혈장 miRNA가 심혈관 위험 예측을 향상할 수 있는지를 평가하고자 하였다. 지

속적으로 혈액투석을 받는 환자에 있어 발병률과 사망률의 주요 원인이 심혈관 질환이라는 사실을 연구의 필요성으로 언급하였다. 연구데이터로는 정기 혈액투석 환자와 건강한 대조군을 대상으로 생존 및 심혈관 발생에 대해 평가한 선행연구 데이터를 활용하였다. 이를 바탕으로 유전자 특성 및 miRNA에 따른 환자-대조군 분류를 위해 기계학습 기반의 'Cox 회귀' 알고리즘을 적용하였다. 또한, miRNA의 질병 유발 영향을 확인하기 위해 회귀 트리 모델을 활용하였다. 회귀 트리 모델로는 기계학습 알고리즘인 'Classification and Regression Tree(CART)'를 사용하였다. 연구결과 순환 miRNA가 심혈관 위험의 바이오마커로써 활용 가능한 유용성 있는 지표라는 사실을 확인하였으며, 심혈관 질환 위험 평가에 있어 회귀 트리 모델의 활용이 적합함을 확인하였다. 궁극적으로 심혈관 질환 발생 위험도에 따라 환자 맞춤형 관리, 치료 및 모니터링을 위한 다양한 전략을 구성할 수 있으며, 임상에서 위험도가 높은 환자를 집중적으로 관리하는 데 유용하게 활용될 수 있다.

Bahado-Singh et al(2022a)은 태반 조직의 후성유전학적 차이와 자폐증 발달과의 연관성 확인하기 위해 기계학습 기술을 활용하고자 하였다. 연구데이터로 신생아 자폐증 환자와 건강한 대조군을 대상으로 태반 조직을 추출하여 활용하였다. 알고리즘으로는 딥러닝(Deep Learning), 서포트 벡터 머신(Support Vector Machine, SVM), 일반화 선형 모델(Generalized Linear Model, GLM), 마이크로어레이를 위한 예측분석(Prediction Analysis for Microarrays, PAM), 랜덤 포레스트(Random Forest, RF), 선형판별분석(Linear Disrciminant Analysis, LDA)을 활용하였다. 이를 통해 자폐군에서 유의하게 차등 메틸화된 CpG에 대해 자폐증과의 연관성을 감지하였다. 태반 표본에 인공지능 분석을 활용하여 미숙아의 자폐증 후속 발병에 대한 예측모델을 수립하고, 자폐증의 분자적 메커니즘을 확인할 수 있었다.

Tran et al(2022)의 연구는 정밀의학을 달성하는 것을 궁극적인 목표로 하여, 암 치료를 위한 병리학적 분류 도구의 효과성을 검증하고자 하였다. 이에 중추신경계 종양 분류에서 메틸화 데이터에 대한 기계학습 적용을 탐구하고자 하였다. 활용 데이터로는 종양 참조 코호트(GSE90496)의 게놈 전체 DNA 메틸화 데이터를 활용하였다. 기계학습 알고리즘은 one nearest neighbor(oneNN), 의사결정나무 C5.0, 서포트 벡터 머신(Support Vector machine, SVM)을 활용하였다. 연구결과 기계학습이 DNA-메틸화 프로파일을 사용하여 분류되지 않은 샘플을 분류해내는데 효율적인 접근 방식임을 확인하였다. 결과적으로 공개 메틸화 데이터를 활용하여 인공지능 분류모델의 성능을 향상시키는 데 도움이 될 수 있으며, 검증 목적으로 근거 자료 레이블을 제공할 수 있는 역량을 가진다고 언급하였다. 궁극적으로 임상 환경에서 인공지능을 활용한 분류 레이블 생성은 정밀의학에서 적절한 암 치료 제공을 위한 도구로 활용될 수 있으며, 전문가의 투입을 줄이는 데 이바지할 수 있음을 언급하였다.

Bahado-Singh et al(2022b)의 연구에서는, 소아 대상 선천성 심장결함에 집중하였다. 심혈관 질환 분야에서는 영상데이터를 분석하며 심부전 등 성인 심혈관 질환을 진단하기 위해 인공지능 기술이 활용되고 있으나, 소아 대상의 선천성 심혈관 결함에는 거의 적용이 이루어지지 않고 있다는 점을 한계로 제시하였다. 또한 대동맥협착(Coarctation of the aorta, 이하 'CoA')는 선천성 심장결함이며, 산전 초음파 및 신생아 맥박산소측정검사를 시행할 경우 CoA의 예측 정확도가 낮은 문제점이 존재한다. CoA의 진단이 늦어질 경우 심각한 합병증을 초래할 우려가 있다는 연구의 필요성에 따라 CoA에 중요한 후성유전학적 변화가 있는지 확인하고자 본 연구를 수행하였다. 이에 인공지능을 활용하여 CoA에의 후성유전학적 변화를 확인하고 관련 유전자를 기반으로 질병발생의 잠재적 예후를 확인하고자 하였다. 환자 데이터는 미시간 보건복지부(Michigan Department of Health & Human Services, 이하 'MDHHS')를 통해 수집하였으며, 연구기간 동안 해당 기관

에 보고된 가장 최근 24건의 CoA 사례로 제한하였다. 대조군 데이터는 MDHHS 데이터베이스에서 무작위로 추출하였다. 랜덤 포레스트(Random Forest, RF), 서포트 벡터 머신(Support Vector Machine, SVM), 선형 판별 분석(Linear Discriminant Analysis, LDA), 마이크로어레이에 대한 예측분석(Prediction analysis for microarrays, PAM), 일반화 선형 모델(Generalized linear Model, GLM)의 기계학습 알고리즘이 활용되었다. 연구결과 후성유전학 데이터 분석을 통해 비증후군 CoA를 예측하였으며, 병원성 증거를 확인하였다. 예측모델은 출산 후 정상 및 비정상적인 심장발달 또는 기능 관련하여 선천성 심혈관 결함을 모니터링하고 잠재적 CoA을 예측하는 데 도움을 줄 수 있었다. 또한, 환자군을 미리 선별하고 모니터링 혹은 치료개입의 우선순위를 확인할 수 있었다. 궁극적으로 연구의 결과는 정밀의학의 심혈관 분야에 중요한 도구로 활용될 잠재성이 존재한다고 언급하였다.

Yu Y et al(2021)의 연구에서는 기계학습 기술을 활용하여 초기 단계인 침윤성 유방암 환자에서 액와림프절 상태를 평가하는 방사선학 평가 접근법을 개발하고자 하였다. 더불어 방사선학과 종양 미생물 환경과의 관련성을 확인하고자 하였다. 랜덤 포레스트(Random Forest, RF), 서포트 벡터 머신(Support Vector Machine, SVM), CIBERSORT 알고리즘이 활용되었다. 사용한 데이터로는 Sun Yat-sen Memorial Hospital of Sun Yat-sen University, Sun Yat-sen University Cancer Center, Shunde Hospital of Southern Medical University, Tungwah Hospital of Sun Yat-sen University의 유방암 환자 모집 데이터, 연구 그룹의 다중 시퀀스 MRI 이미지는 Picture Archiving and Communication System에서 검색되었으며, 3D Slicer 소프트웨어의 N4ITK Bias Field Correction 모듈을 적용하였다. 연구결과 초기 침습성 유방암에서 액와림프절 전이 환자를 확인할 수 있는 다중오믹 서명을 확인할 수 있었다. 다중오믹 서명을 바탕으로 임상 및 병리학적 상태가 일치하지 않는 환자에 있어서 액와림프절 상태 평가의 정확도를 향

상시킬 수 있었다. 결과적으로 다중오믹 서명이 액와림프절 상태의 예측에 효과적인 도구임을 나타내며, 수술 전 진단 및 치료 결정에 유용한 정보를 제공할 수 있다는 함의점이 도출되었다. 유방암 환자는 초기 진단 시에 액와림프절의 상태에 따라 다른 예후를 보이며, 그에 따라 치료 계획도 상이하게 제공하여야 할 필요성이 있다. 예를 들어 액와림프절 상태가 양성인 환자는 음성인 환자보다 예후가 나쁘다는 것이 입증되어 있어, 양성상태를 가진 환자는 가이드라인에 따라 고위험 환자로 간주되어 보충요법을 받아야 한다. 이에 액와림프절 상태에 대한 오진은 환자의 과다 치료와 의료비용 낭비의 원인이 될 수 있다는 문제가 있다. 이러한 측면에서 연구를 통해 개발된 인공지능 모델은 치료 경비를 줄이고 더 나은 치료를 제공하는 데 도움을 줄 수 있으며, 진단 예측 및 치료의 효과성을 향상하는 데 도움이 될 수 있음을 언급하였다.

Diboun et al(2021)은 DNA 메틸화 프로파일의 '뼈의 파젯병' 관련 변형에 의한 역할을 확인하고자 하였다. 기계학습 기반의 선형회귀모델(generalized linear model) 알고리즘을 활용하여, '뼈의 파젯병' 환자와 대조군 간 DNA 메틸화 프로파일을 비교하였다. 데이터는 Infinium HumanMethylation450 BeadChip(Illumina, USA)를 활용하였다. 연구결과 대조군 대비 '뼈의 파젯병' 환자에서 DNA 메틸화 프로파일 특성이 발견되었다. 골세포 분화와 같은 뼈 관련 기능을 포함하여 '뼈의 파젯병'의 병리발생과 기능적 관련성이 있는 유전자 내 또는 근처에 위치하고 있었다. 결론적으로 후성유전학적 요인이 '뼈의 파젯병' 발병에 기여한다는 사실이 밝혀졌다. 연구결과는 궁극적으로 질병의 예측을 위한 진단 지표로서 DNA 메틸화 프로파일이 활용할 수 있다는 사실을 밝혀내었다. 이 연구는 후성유전학적 표지자가 유전자 프로파일링과 결합할 경우 임상적으로 적절한 경우 조기 개입을 고려할 수 있도록 질병 가족력이 있는 사람들의 '뼈의 파젯병' 발생 위험을 평가하는 수단으로 활용할 수 있는 가치가 있음을 언급하였다.

Karisola et al(2021)의 연구에서는 달걀을 이용한 경구 면역요법을 통해 탈감작을 유도하는 세포 매개 분자 메커니즘을 확인하는 것을 연구의 목표로 하였다. 즉, 면역요법 과정에서 유전자 발현 변화가 알레르기 반응(체액 반응 및 임상 결과)과 상관관계가 있는지 확인하고자 하였다. 연구를 위해 CIBERSORT 알고리즘을 이용하여 검증된 유전자 서명 행렬을 이용, 22개의 개별 면역세포의 백혈구 부분 집합 비율을 추정하였다. 라벨링 및 배열 분산에서 발생하는 배치 효과를 제거하기 위해 Combat 기능이 있는 Surrogate Variable Analysis(SVA) 패키지를 사용하였다. 유클리드 방법과 k-means 알고리즘은 각각 두 데이터셋 포인트 간의 거리를 결정하고 두 클러스터 간의 거리를 정의하는 데 사용되었다. 활용 데이터는 NCBI의 Gene Expression Omnibus(GEO) Database 내의 데이터를 사용하였다. 면역요법 0개월, 3개월, 8개월 후 유전체 전반의 유전자 발현 변화가 체액 반응 및 임상 결과와 상관관계가 있는지 확인하고자 하였다. 연구결과, 달걀을 이용한 경구 면역요법을 시행하는 동안 알레르겐 탈감작에 관여하는 주요 유전자, 생물학적 과정 및 세포 유형을 식별하기 위해 체액 매개자(항체 및 사이토카인)의 결과와 고급 유전자 발현 분석을 통합할 수 있었다. 결과적으로 연구는 달걀을 이용한 경구 면역요법의 결과를 예측하기 위한 잠재적인 바이오마커를 개발할 수 있게 하였다. 개별적으로 조정되고 개인화된 치료 프로토콜의 수립에 이바지할 수 있다는 함의점을 도출하였다.

Aref-Eshghi et al(2018)의 연구는 염색질 조절 단백질의 상호 연관된 기능을 고려하여, 표준 임상 진단을 보완하기 위해 증후군 특이적 바이오마커를 제공할 수 있는 DNA 메틸화 변이를 확인하고자 하였다. 소아발달장애, 후성유전기계, 가부키증후군, CHARGE 증후군, 플로팅하버증후군, ATRX증후군, 소토스증후군, 클라에스젠센증후군에 초점을 맞추었다. 활용 데이터는 Care4Rare Canada Consortium에서 ADCA-DN, Copin-Siris 증후군, SBBYSS, GTPTS 및 Floating-Harbor 증후군을 가진 피험자의 말초혈액 DNA 샘플을 수집하였다. 또한 Greenwood(SC, USA), NCBI

Gene Expression Omnibus(GEO) Database에서 가부키와 CHARGE 증후군 코호트를 수집하였다. 수집한 데이터를 바탕으로 서포트 벡터 머신(Support Vector Machine, SVM)을 활용하여 다중계층 예측모델을 구축, 질병과 관련된 유전자를 가진 환자를 분류하고자 하였다. 이에 14개의 멘델계 장애를 포괄하는 대규모 개체 집단의 말초혈액 샘플을 조사하였다. 연구결과, 부분적으로 중복되는 DNA 메틸화 서명이 질병 유전자의 변이 유발과 관련이 있음을 확인하였다. 이러한 DNA 메틸화 변이를 결합하여 민감도와 특이성이 높은 여러 증후군을 동시에 선별할 수 있는 기계학습 도구의 개발이 가능함을 입증하였다.

3.2.3.2. 환자 분류를 위해 인공지능을 적용한 연구

Shokhirev & Johnson(2022)에서는 기계학습 기술을 바탕으로 알츠하이머 질병의 예측모델을 구축하여 알츠하이머(Alzheimer's Disease, 이하 'AD') 환자와 AD 환자가 아닌 대조군을 구분하고자 하였다. 생물학적 데이터로 RNA-Seq, 마이크로어레이, 프로테오믹스 및 마이크로 RNA 표본을 수집하고 메타분석을 수행함으로써 대규모 다중 Omics 데이터 세트를 생성하였다. 이후 기계학습 기반의 '랜덤 포레스트(Random Forest, RF)', bagFDA, 일반화된 정규화 선형 모델(generalized regularized linear model, glmnet) 알고리즘을 활용한 예측모델을 통해 생물학적 데이터 샘플을 AD 질병군으로 명확하게 분류할 수 있는 유전자와 단백질을 식별하고자 하였다. 네트워크 토폴로지 기반 분석을 통해 상위 예측 유전자와 단백질을 분석하고 Gene Ontology Biological Process 데이터베이스를 교차 참조하였다. 상위 예측 miRNA에 대해서는 별도의 접근법을 활용하여, 연령대에서 AD 진단을 하는데 가장 중요한 10개의 miRNA를 miRwalk을 이용하여 이들의 예측 유전자 타겟을 파악하기 위하여 분석하였다. 이러한 예측 유전자들은 Reactome 데이터베이스를 활용, 유전자 세트 농축 분석을 수행하였다. 연구결과 AD 질병군을 가진 환자는 연령에 따라 후성유전학적 변이를 유발하는 관련 유전

자, 단백질 및 miRNA의 변화가 상이한 결과를 보였다. 즉 나이 예측과 관련된 유전자와 단백질이 인간의 수명 또는 연령 관련 질병에 영향을 미치는 결과를 도출하였다. 이에 연구에서 구축된 예측모델을 활용하여 비교적 낮은 연령대에서 대표되는 질병 유발 표적 특징과 비교적 높은 연령대에서 대표되는 표적 특징을 비교 가능함으로써, 연령 맞춤형 임상적 의료개입의 시도 가능성을 언급하였다.

Hess et al(2020)에서는 양극성 장애(Bipolar Disorder, BD)를 가진 환자와 건강한 대조군을 대상으로 말초 혈액에서 전사체 유전자 발현 데이터를 확보하여, 주요 정신질환(BD, 조현병)의 표적 특징과 병원성 기전을 확인하고자 하였다. 유전자의 표현을 종속변수로, 진단을 독립변수로 지정한 다변량 선형회귀모델을 활용하여 유전자 당 질병 발현 추정치를 분석하였다. 이후 유전자 당 질병 발현이 많이 발생하는 수준을 기반으로 연구 대상자의 진단상태 예측을 위해 기계학습 기술 기반의 모델을 구축하였다. 활용 알고리즘으로는 '랜덤 포레스트(Random forest, RF)', 'Linear 커널 서포트 벡터 머신(Linear-kernel support vector machines, SVM)', '방사형 커널 서포트 벡터머신(radial-kernel, SVM)'과 같은 다중기계학습 알고리즘을 적용하였다. 연구결과 BD를 가진 환자군에서 유의하게 차별적으로 발현되는 유전자를 확인하였다. 결론적으로 주요 정신질환(BD, 조현병)을 앓는 환자의 혈액에서 특정 유전자와 유전자 세트의 특성 발현을 확인하였으며, 이러한 연구결과를 바탕으로 후성유전학적 표적 특징 조절을 통해 주요 정신질환의 발생을 완화할 수 있다고 언급하였다.

Kalyakulina et al(2022)에서는 파킨슨병과 조현병을 예시로 하여 대규모 DNA 메틸화 데이터를 기반으로 한 환자-대조군 분류를 목적으로 워크플로우를 개발하고자 하였다. 파킨슨병이나 조현병 환자의 DNA 메틸화 데이터 집합을 분석하여, 환자군과 건강한 대조군 분류를 위해 기계학습 알고리즘 중 '서포트 벡터 머신(Support Vector Machine, SVM)', 'XGBoost', 'LightGBM', 'CatBoost'을 활

용하였다. 연구에서 사용된 데이터는 NCBI의 Gene Expression Omnibus(GEO) Database(GSE145361, GSE111629, GSE72774, GSE84727, GSE80417, GSE152027, GSE116379)를 통해 수집하였다. 연구결과 DNA 메틸화 데이터를 기반으로 환자군을 분류하는 타당한 인공지능 기반 모델을 구축하였다. 또한, 기계학습 알고리즘 모델에 훈련하는 표본의 크기가 클수록 모델의 품질이 향상된다는 사실을 확인하였으며, 설명 가능한 인공지능 알고리즘 기반의 예측모델 구축을 기반으로 분류 정확도를 향상함으로써 인구집단 및 환자 개인 중심의 관점 모두에서의 질병 예방이 가능하다는 시사점을 제시하였다.

Bendifallah et al(2022)의 연구는 자궁내막증을 유발하는 유전자를 검토하고 자궁내막증 환자와 건강한 대조군 분류를 목적으로 기계학습 기술을 적용하고자 하였다. 연구데이터는 자궁내막증과 연관된 만성 골반통을 앓고 있는 여성으로부터 얻은 혈장 표본을 포함하였다(전향적 ENDO-miRNA 연구). 혈장 표본을 분리하여 RNA 검체를 추출하였다. 추출한 RNA 검체의 miRNA를 점수화하기 위해 기계학습 기반 진단모델을 개발하고자 하였으며, '로지스틱 회귀분석 (Logistic Regression, LR)', '랜덤 포레스트(Random Forest, RF)', 'eXtreme Gradient Boosting(XGBoost)', 'AdaBoost' 알고리즘을 활용하였다. 연구결과 만성 골반통을 앓고 있는 환자군에서 자궁내막증을 유발하는 특정 유전자를 확인하였다. 특정 유전자를 기반으로 환자를 분류하는 인공지능 모델의 활용을 통해 자궁내막증 진단 및 치료시간의 최소화, 치료경과 개선 등에 효과적으로 작용될 수 있다고 언급하였다.

Arabyarmohammadi et al(2022) 연구에서는 골수이형성증후군(Myelodysplastic syndromes, 이하 'MDSs')과 급성 골수성 백혈병(acute myeloid leukemia, 이하 'AML')이 발병률, 사망률 및 재발률이 높으며 치료가 어려운 혈액질환임에 따라 예방 및 예측의 차원에서 골수아세포를 중심으로 생존 예측을 확인하고자 하

였다. MDSs는 말초혈액 저혈구증을 특징으로 하는 혈액이상증후군이며, 급성 골수성 백혈병으로의 전환 위험도가 높은 질병이다. 이 질병들은 혈액과 골수에서 건강한 줄기세포를 해치며, 결과적으로 급성 골수성 백혈병 환자는 감염, 빈혈, 혈액 응고 장애를 경험할 수 있다. 이에 골수아세포에 대한 모니터링 및 진단이 요구된다. 연구자들은 골수아세포의 특징을 사용하여 MDSs 및 AML 환자의 동종조혈모세포이식(Hematopeoietic stem-ell trasplant, 이하 'HCT') 이후 재발 예측을 위한 모델을 구축하였다. 기계학습 기술을 바탕으로 하였으며, 생존 예측을 위해 '라쏘 회귀(Lass Regression)'를 활용하였다. 골수아세포 백분율만을 활용하여 재발 예측하기 위한 알고리즘으로 기계학습 기반의 '선형 판별 분석(Linear discriminant analysis)'을 사용하였다. 활용 데이터로는 훈련 세트(S = 52) 및 검증 세트(S = 40)에 무작위로 할당된 MDSs나 AML 환자 92명으로부터 Wright-Giemsa 염색 후 HCT 흡인산염 이미지를 수집하였다. 분석 결과 골수아세포의 크로마틴 변형 패턴을 확인하여 MDSs나 AML의 재발을 예측할 수 있으며, 골수아세포의 염색질 질감에 따라 AML 재발이 가능한 환자를 예측할 수 있음이 확인되었다. 다만 연구에 활용된 샘플의 검증 뿐 아니라 대규모 전향적 임상시험 평가를 바탕으로 추가적인 검증 필요성을 언급하였다. 이러한 연구결과를 통해 궁극적으로 질병의 재발을 예측하고 환자 개인에게 적합한 예방적 차원의 의료개입을 가능하게 할 수 있다.

Bahado-Singh et al(2023)의 연구에서는 태아 선천성 심장결함의 최소 침습적 검출을 위해 유전체 전체 DNA 메틸화와 순환 세포가 없는 DNA에의 기계학습 분석을 구현하는 것을 목표로 하였다. DNA 메틸화는 후성유전학 측면에서 심장발달의 유전자 발현을 조절하는 중요한 메커니즘으로 작용한다. 이에 산모 혈액 내 순환 세포가 없는 DNA 전체 유전체에 대한 후성유전학적 분석의 결합을 통해 태아 선천성 심장결함을 예측하고자 하였다. 이를 위한 기계학습 기반 알고리즘으로는 '랜덤 포레스트(Random forest, RF)', '서포트 벡터 머신(Support Vector

Machine. SVM)', '마이크로어레이 예측분석(prediction analysis for microarrays)', '일반화 선형모형(generalized linear model)'을 활용하였다. 환자 데이터는 임신 2, 3기에 분리된 태아 선천성 심장결함에 대한 산전초음파 의심 또는 진단에 근거하여 전향적으로 수집하였다. 예측모델 생성을 위해 선천성 심장결함 환자군과 건강한 대조군의 데이터를 기반으로 선천성 심장결함 환자군을 식별하기 위해 기계학습 알고리즘을 훈련했다. 연구결과 선천성 심장결함 환자군에서 상당한 DNA 메틸화 변화를 확인할 수 있었다. 즉 후성유전학적 메커니즘은 유전자 발현을 조절하여 태아의 심장 생성에 영향을 미치는 것으로 확인되며, 태반 DNA의 메틸화 변화가 선천성 심장결함을 유발할 수 있다. 장기적으로 이러한 연구결과는 태반 혈액검사와 같은 최소 침습적 모니터링을 수행하여 산전검사의 효과를 높이는 데 도움을 줄 수 있다. 또한, 심혈관 질환의 발병 과정을 이해하고 환자 개인에게 적합한 정밀 표적 치료 및 예방 전략 수립 등 임상적 차원의 해결책을 제시하는 데 유용한 참고자료로 활용될 수 있다.

3.2.3.3. 치료의 예후를 확인하기 위해 인공지능을 적용한 연구

Lin et al(2020)의 연구에서는 알코올 의존증 치료 관련하여 날트렉손(Naltrexone, 이하 'NTX') 치료제를 활용 후 알코올 의존증의 재발 관련하여 DNA 메틸화의 영향을 확인, 재발 위험을 예측하고자 하였다. NTX는 알코올의 강화 효과를 감소시키는 약물로, 날트렉손 치료제의 효과는 환자의 유전적 배경, 음주 상황, 니코틴의 동반 사용과 같은 흡연 상태 및 기타 요인에 따라 다르다. 이에 유전적 요인과 환경적 요인(만성 알코올 소비 포함) 모두 DNA 메틸화 상태에 영향을 미치는지 확인하고, NTX 치료 후 DNA 메틸화가 알코올 의존증의 재발에 미치는 영향을 확인하고자 하였다. 연구는 93명의 알코올 의존증 질환자를 선정하였으며, 알코올 의존증에 있어 NTX 치료 결과에 대한 수용체 유전자 변화가 있는 참가자 중에서 선정되었다. 알코올 의존증 재발 예측을 위해 기계학습 기술

중 ‘랜덤 포레스트(Random forest, RF)’ 알고리즘을 활용하였으며, 알코올 의존증 재발 위험에 미치는 영향을 확인하였다. 베이지안 분석을 통해 인종, 나이, 치료제 종류 중 나이가 알코올 의존증 재발 위험에 영향을 미치는 요인으로 도출되었다. 연구결과 나이가 많을수록 알코올 의존증 재발 가능성이 낮아진다는 사실을 확인하였지만, 개별 단위의 DNA 메틸화 변화가 NTX 치료 후 재발에 미치는 유의한 영향은 검증되지 않았다. 궁극적으로 이러한 연구결과는 DNA 메틸화 변화가 인간의 높은 나이와 높은 상관관계가 있음을 암시한다고 볼 수 있다. 즉, 나이가 보상 또는 중독과 관련된 유전자의 후성유전학적 상태에 미치는 영향을 반영하며, 이로 인해 높은 나이의 환자들이 약물치료에 더 반응적일 수 있다고 할 수 있다. 또한 연령과 유전적 변이 간의 상호작용이 NTX 치료 효과에 영향을 줄 수 있는 것이다. 연구자들은 DNA 메틸화 변화가 질병의 재발에 직접적으로 미치는 영향은 확인되지 않았으나, DNA 메틸화 변화 또는 유전변이의 효과가 어떻게 발생하는 지에 대한 메커니즘의 탐구가 요구된다고 언급하였다. 알코올 의존증 치료효과를 향상하기 위해서는 알코올 의존증이 재발할 가능성이 낮은 환자들의 무의미한 약물 노출을 피하는데 도움이 될 수 있으며, 환자의 후성유전학적 상태를 바탕으로 치료에 대한 효과를 최적화하는 특정 약물 치료법을 활용함으로써 정밀의료개입의 실현을 확대할 수 있다고 제언하였다.

Huang et al(2021)은 증가하는 보조생식술(assisted reproductive technology, 이하 ‘ART’)사용 대비 출생한 사람의 장기적인 건강 상태에 대해 연구한 사실이 없다는 한계를 제시하며, ART를 통해 출생한 사람의 장기 심혈관 위험에 대해 평가하였다. 메틸화 데이터 및 기본 참가자 특성은 NCBI Gene Expression Omnibus(GEO) Database에 시리즈 등록 번호 GSE158064를 통해 확인할 수 있으며, 익명화된 데이터셋과 원시 결과는 github(github.com/jhuang35/ivf_growth/) 및 Zenodo(10.5281/zenodo)를 통해 게시하였다. 기계학습 기술 중 ‘XG Boost(boosted trees)’ 알고리즘을 활용하였으며, ART를 통한 임신과 아동의 키

및 혈압 간의 관계를 확인하기 위해 DNA 메틸화의 차이가 있는지 확인하고자 하였다. 연구결과, ART를 통해 출생한 사람과 자연임신 후 출생한 사람의 DNA 메틸화에는 낮은 연관성이 있었다. 그러나 DNA 메틸화와는 무관하게 ART를 통한 임신 후 출생아동의 키와 몸무게를 감소시키는 것과 관련이 있었으며, 피하지방의 두께, 지방의 질량 및 혈압도 낮은 것으로 확인되었다. 반면 아동의 연령이 증가할수록 ART를 통해 출산한 아동과 자연임신 출산 아동 간 혈압은 유의미한 차이를 보이지 않았다. 이는 아동연령이 어릴 때 관찰된 혈압 수치의 차이가 부모의 유전적 영향에 의한 차이나 생식능력의 차이로 인해 발생한 것이라고 판단할 수 있다. ART를 통해 임신한 부모들은 아동의 건강을 위해 신체활동, 체중 감량, 금연 등의 건강행동을 유지하는 등 자연임신 부모에 비해 복잡한 선택에 노출되고, 고령의 나이일 수 있으며, 기질적 내분비·비만·다낭성 난소 증후군과 같은 대사질환을 가질 수 있다는 사실을 시사하였다.

3.2.4. 결과

주제범위 문헌고찰에 포함된 연구들에서 가장 많이 적용된 인공지능 기술은 '기계학습(Machine Learning)'이며, 기계학습의 하위분류 알고리즘 중에는 '랜덤 포레스트(Random forest, RF)'과 '서포트 벡터 머신(Support Vector Machine, SVM)'이 가장 많이 활용되고 있었다.

대부분의 연구에서 후성유전학적 발생기전인 'DNA 메틸화'와 관련된 패턴 특이성이나 유전자 발현특성, DNA 메틸화와 질병 또는 치료 예후 간 상관성 등에 분석의 초점을 두고 있었다. 관련 질병으로는 암 및 심혈관 질환과 관련된 연구가 9편으로 가장 많았다. 정신질환 관련 연구 5편, 감염성 질환 관련 연구 2편 순으로 많았다. 이외에도 유전, 면역 및 알레르기 질환 관련 연구, 뇌질환, 부인과 질환, 혈액질환, 만성질환, 골질환 관련 연구가 확인되었다.

또한, 연구들에서 활용하고 있는 데이터 세트는 각 병원에서 수집된 환자 데이터가 대부분이었다. 수집한 환자 데이터는 NCBI의 GEO Database에 저장 및 게시되기도 하였다. 이외 연구에서는 GEO Database를 통해 기존에 게시되어 있던 여러 자료를 연구 데이터 세트로 설정하여 활용하고 있었다.

앞서 설정한 연구문제에 의거, 주제범위 문헌고찰의 요약적 결과를 다음과 같이 종합 정리하였다(표 3).

표 3. 주제범위 문헌고찰 결과

번호	저자	사용된 인공지능 기술;모델	인공지능 수행 작업 유형	연구초점	관련질병; 후성유전 발생기전	데이터 세트
1	Imgenberg-Kreuz et al (2019)	기계학습; 랜덤포레스트	전신홍반성 루푸스, 일차 쇼그렌 증후군의 질병특이적 변화감지	자가면역질환에서 유전자 전체 DNA 메틸화 패턴 중 관련 특성 추출 기계학습 능력 검증	전신홍반루푸스, 일차쇼그렌 증후군; DNA 메틸화	Interferome v2.01 database, ToppGene Suite
2	J Orozco et al (2018)	기계학습; 랜덤포레스트	뇌종양 유형별 DNA 메틸화 특성 확인, 뇌종양 식별	뇌종양의 최적 치료를 목적으로 뇌종양 유형에 따른 후성유전 특징 확인 및 뇌전이 분류	뇌종양; DNA 메틸화	의료센터 환자데이터, NCBI Gene Expression Omnibus (GEO) Database
3	ElHefnawi et al (2013)	기계학습; 서포트 벡터 머신	간세포암종의 차별적발현 miRNA 표적 유전자 예측 및 분석	간세포암종 환자 예후와 생존율 향상을 위해 조기 진단을 통한 약물표적, 치료 개입 전략 수립	간세포암종; DNA 메틸화	PhenomiR database (www.mips.helmholtz-muenchen.de/phenomir/)
4	Min et al (2018)	기계학습; 서포트 벡터 머신	수족구병과 관련된 주요 표적 특징 도출	수족구진단, 바이오마커 이용 감염병 진단 워크플로우 개발	수족구병; DNA 메틸화	Kandang Kerbau(KK) 여성아동병원 환자데이터
5	Yang et al (2018)	기계학습; 서포트 벡터 머신	흡연 관련 특징유전자 식별 도구 유효성 검증	폐선암의 유전적 메커니즘 확인	폐선암; DNA 메틸화	NCBI Gene Expression Omnibus (GEO) Database

표 3. 주제범위 문헌고찰 결과(계속)

번호	저자	사용된 인공지능 기술;모델	인공지능 수행 작업 유형	연구초점	관련질병; 후성유전 발생기전	데이터 세트
6	Chakravarthy et al (2016)	기계학습; K-최근접 이웃, 랜덤 포레스트	HPV의 암종 인과적 역할 확인	분류자를 사용, 암종에의 HPV 인과성 예측	인유두종바이러스; DNA 메틸화	ArrayExpress, TCGA Data Portal, Cancer Browser
7	Huan T et al (2019)	기계학습; Elastic-cox ph, 랜덤 생존 포레스트, Cox-nnet, DeepSurv	DNA 메틸화 데이터 기반 심혈관 질환 사망률 및 암 사망률 예측	사망 예측을 위한 정보 바이오마커 유용성 확인	암, 심혈관질환 사망; DNA 메틸화	추적관찰 (10년) 환자데이터
8	Cheng et al (2023)	기계학습; 라쏘 회귀, stepwise COX	새로운 전립선암 예측 시그니처 구성 및 검증	의료개입 및 정밀성향상, 전립선암 예후 및 치료반응 예측	전립선암; non-coding RNA (ncRNA)	The Cancer Genome Atlas (TCGA) 코호트
9	de Gonzalo-Calvo et al (2020)	기계학습; Cox 회귀, 회귀트리모형(CART)	혈액투석 말기신장질환자 심혈관 위험 프로파일 식별 miRNA 유용성 확인	잠재적 심혈관 위험 가능성에 대한 예측, 위험 평가	말기신장 및 심혈관질환; DNA 메틸화	선행연구 데이터 활용
10	Bahado-Singh et al (2022a)	기계학습; 딥러닝, 서포트벡터머신, 일반화선형모델, 마이크로어레이예측분석, 랜덤포레스트, 선형판별분석	자폐군에서 유의하게 차등 메틸화된 CpG에 대해 자폐증 감지	미숙아 자폐증 발병 예측모델수립, 자폐증 분자적 메커니즘 확인	자폐증; DNA 메틸화	환자 데이터 수집

표 3. 주제범위 문헌고찰 결과(계속)

번호	저자	사용된 인공지능 기술;모델	인공지능 수행 작업 유형	연구초점	관련질병; 후성유전 발생기전	데이터 세트
11	Tran et al (2022)	기계학습; one nearest neighbor, Decision Tree C5.0, 서포트 벡터머신	DNA-메틸화 프로파일을 사용하여 분류되지 않은 샘플분류	중추신경계 종양분류 메틸화데이터 기계학습 적용 탐구	중추신경계 종양; DNA 메틸화	NCBI Gene Expression Omnibus (GEO) Database
12	Bahado-Singh et al (2022b)	기계학습; 랜덤포레스트, 서포트 벡터머신, 선형판별분석, 마이크로어레이 예측분석, 일반화 선형 모델	비증후군 대동맥 협착 예측, 병원성 증거 확인	대동맥협착의 후성유전학적 변화확인, 질병발생의 잠재적 예후확인	신생아 대동맥 협착증; DNA 메틸화	미시간 보건복지부 (MDHHS) 데이터 베이스
13	Yu Y et al (2021)	기계학습; 랜덤포레스트, 서포트 벡터머신, CIBERSORT 알고리즘	초기 침습성 유방암에서 액와림프절 전이환자 분류를 위한 시그니처 확인	초기 단계인 침윤성 유방암 환자에서 액와림프절 상태 평가	초기 침습성 유방암; ncRNA, DNA 메틸화	유관병원 환자데이터, Picture Archiving and Communication System, N4ITK Bias Field Correction Module
14	Diboun et al (2021)	기계학습; 선형회귀모델 (generalized linear model)	뼈의 파젯병에서 DNA 메틸화 프로파일 확인	가족력 있는 사람들 뼈의 파젯병 발생 위험 예측, 예방수단 개발	뼈의 파젯병; DNA 메틸화	Infinium HumanMethylation450 BeadChip (Illumina, USA)
15	Karisola et al (2021)	기계학습; CIBERSORT 알고리즘, k-means 알고리즘	계란 경구 면역요법을 통한 세포 매개 분자 메커니즘 확인	면역요법의 알레르기 염증, 체액 반응 조절 예측	계란 알레르기; ncRNA	NCBI Gene Expression Omnibus (GEO) Database

표 3. 주제범위 문헌고찰 결과(계속)

번호	저자	사용된 인공지능 기술;모델	인공지능 수행 작업 유형	연구초점	관련질병; 후성유전 발생기전	데이터 세트
16	Aref-Eshghi et al (2018)	기계학습; 서포트 벡터머신,	증후군 특이적 바이오마커인 DNA 메틸화 에피-시그니처확인	민감도와 특이성이 높은 여러 증후군을 동시에 선별할 수 있는 도구 개발	유전자 돌연변이 증후군; DNA 메틸화	Care4Rare Canada Consortium, Greenwood(SC, USA), NCBI Gene Expression Omnibus(GEO)
17	Shokhirev & Johnson (2022)	기계학습; 랜덤포레스트, bagFDA, 일반화된 정규화 선형 모델	알츠하이머 예측모델 구축, 환자 구분	알츠하이머 질병군 분류	알츠하이머; DNA 메틸화	Gene Ontology Biological Process, Reactome
18	Hess et al (2020)	기계학습; 랜덤포레스트, 방사형 커널 서포트 벡터 머신	양극성 장애 구별	면역 신호 유전자 발현. 혈액기반의 유전자 발현 기계학습 분류모델구축	정신질환; 히스톤 변형	환자 데이터
19	Kalyakulina et al (2022)	기계학습; 서포트 벡터머신, XGBoost, LightGBM, CatBoost	파킨슨병이나 조현병 환자의 DNA 메틸화 데이터 집합 분석 및 환자군 분류	환자군을 분류 인공지능 기반 모델 구축	파킨슨병, 조현병; DNA 메틸화	NCBI Gene Expression Omnibus (GEO) Database
20	Bendifallah et al (2022)	기계학습; 로지스틱 회귀분석, 랜덤 포레스트, eXtreme Gradient Boosting,AdaBoost	자궁내막증 유발 유전자 검토, 환자 분류	자궁내막증을 유발하는 miRNA 검토, 혈액기반 진단 시그니처 개발	자궁내막증; DNA 메틸화	환자데이터

표 3. 주제범위 문헌고찰 결과(계속)

번호	저자	사용된 인공지능 기술;모델	인공지능 수행 작업 유형	연구초점	관련질병; 후성유전 발생기전	데이터세트
21	Arabyarmohammadi et al (2022)	기계학습; 라쏘 회귀, 선형 판별 분석	골수아세포의 크로마틴 변형 패턴을 확인, 골수이형성증후군이나 급성 골수성 백혈병 재발예측	골수아세포 생존 예측 확인	골수이형성증후군, 급성 골수성 백혈병; 크로마틴 변형	환자데이터
22	Bahado-Singh et al (2023)	기계학습; 랜덤포레스트, 서포트 벡터머신, 마이크로어레이 예측분석, 일반화 선형모형	선천성 심장결함 환자군 식별	태아선천성 심장결함의 최소 침습적 검출	태아 선천성 심장결함; DNA 메틸화	환자데이터
23	Lin et al (2020)	기계학습; 랜덤포레스트	알코올 의존증 재발의 예측 변수로 사용될 입력 변수 그룹 식별	날트렉손 치료 후 알코올 의존증 재발 관련 DNA 메틸화 영향 확인	알코올 의존증; DNA 메틸화	환자데이터
24	Huang et al (2021)	기계학습; XG Boost (boosted trees)	보조생식술 아동의 키 및 혈압 간의 관계 확인을 위해 DNA 메틸화의 차이 확인	보조생식술을 통해 출생한 사람의 장기적 건강 상태 연구	여러 유형의 질병(혈압, 비만 등); DNA 메틸화	NCBI Gene Expression Omnibus (GEO) Database

3.3. 소결

주제범위 문헌고찰에 포함된 연구에서는 각 병원의 환자 데이터, NCBI의 GEO Database, TCGA와 같은 대규모의 데이터 저장소에 게시된 데이터를 활용하고 있었다. 그러나 후성유전적 발생기전을 유발하는 주요요인인 개인 생활습관, 환경요인 등의 데이터는 함께 고려되지 않고 있었다.

인공지능 기술은 주로 다음의 목적으로 적용되고 있었다. 첫째, '질병 예측을 위해 인공지능을 적용한 연구' 에서는 인공지능 기반의 예측모델을 구축하여 질병을 유발하는 특정 발현 유전자를 추출하였다. 이러한 연구결과는 발현 유전자를 가진 사람을 대상으로 질병발생을 최소화하기 위한 예방적 의료개입을 제공하는 데 활용될 수 있다. 둘째, '환자군 분류를 위해 인공지능을 적용한 연구' 에서는 데이터 표본을 바탕으로 후성유전학적 변이를 유발하는 요인을 확인하였다. 이러한 연구결과는 대상 인구집단에서 질병 유발 표적 특징 유전자를 가진 환자를 추출하여 진단 및 치료시간 최소화, 예후 개선에 도움이 될 수 있다. 셋째, '치료의 차별적 예후 확인을 위해 인공지능을 적용한 연구' 에서는 치료개입 결과에 후성유전학적 특성이 미치는 영향을 확인하고자 하였다. 특히 Lin et al(2020)에서는 유전요인과 환경요인을 모두 고려하여 DNA 메틸화 변이를 유발하는 요인을 확인하였다. 후성유전학적 상태를 바탕으로 치료에 대한 효과를 최적화하는 특정 치료법을 활용함으로써 정밀의료개입의 실현을 확대하고자 하였다.

종합적으로 후성유전 관련하여 인공지능을 구현한 연구에서는 후성유전학적 발생기전과 질병 사이의 관계성을 확인하고자 하였으며, 질병 발생 예측에 따라 개개인에게 적합한 의료개입을 가능하게 하는 것을 향후 목표로 설정하였다. 이러한 목표는 유전자-환경 상호작용 및 후성유전 변화에 대한 이해를 바탕으로 환자에 대한 정밀의료 제공에 도움을 줄 수 있다.

제4장 후성유전 분야에서의 인공지능 적용 미래 방향성

4.1. 개요

최근 몇 년 동안 인공지능은 의료분야에서 임상연구, 진단예측 등 다양하게 활용되고 있다(Habuza et al., 2021). 의료분야 인공지능에 대한 투자는 지속적으로 성장하고, 2026년까지 미국 의료경제에 연간 1,500억 달러의 절감 효과를 가져올 것으로 예측된다(Accenture, 2019). 이러한 변화는 정밀의료 개입 및 진단·의료의 운영 및 관리에 영향을 미칠 것이며, 환자 맞춤형 데이터 기반의 의료서비스를 제공할 것으로 기대된다(Insights, 2019; Rauschert et al., 2020). 이는 앞서 진행한 주제범위 문헌고찰 결과를 통해서도 확인할 수 있었다. 인공지능을 통해 통합·분석된 고품질 데이터는 임상 현장에서 의사결정을 지원하기 위한 정확한 예측정보를 제공할 수 있었다(Jaffe et al., 2012). 인공지능 모델을 훈련함으로써 질병 바이오마커를 식별, 개인의 질병 이해도를 높일 수 있었다(Sobia Raza, 2020).

그러나 기존의 환자 데이터나 대규모 데이터베이스의 유전체 데이터를 주요 학습데이터로 사용하고 있는 점, 환경이나 생활습관 데이터 간 상호작용을 고려하지 못하였다는 점에서 한계가 존재한다. 후성유전학적 변화를 유발하여 질병을 발생시키는 유전적 요인 이외에도 유전자-환경 상호작용 및 후성유전 변화에 대한 이해가 필요하다. 통합적 이해를 바탕으로 개개인에 대한 맞춤형 데이터 분석이 이루어져야 한다. 맞춤형 데이터 분석은 맞춤형 의료서비스의 제공, 개인의 만족도 증대로 이어질 수 있다. 궁극적으로 후성유전 분야 연구에서 최종 목표로 설정하고 있는 정밀의료의 실현이 가능하다. 이에 후성유전 측면에서 정밀의료의 실현을 위해 인공지능이 나아가야 할 미래 방향성을 고민해보고자 한다.

4.2. 미래 방향성 제시를 위한 전통적 문헌고찰 방법 적용

4.2.1. 개요

본 연구는 앞서 설정한 세 번째 연구문제 '역학 · 유전학 · 후성유전 각 분야에서 인공지능 기술이 어느 정도 영역까지 적용되어 있으며 후성유전 분야에서의 적용 한계점은 무엇인가?' 에 의거 전통적 문헌고찰 방법을 활용하여 후성유전 분야에서의 인공지능 적용 미래 방향성을 제시하고자 하였다. 선행연구 및 기타 자료를 중심으로 역학, 유전학, 후성유전학으로 구분하여 각 분야에서 인공지능 기술을 어느 정도 영역까지 적용하고 있는지 탐색 및 비교하고자 하였다. 인공지능 기술 영역은 앞서 기술한 '인지', '학습', '추론', '행동' 으로 구분하였다.

이외에도 후성유전 분야에서 활용 가능한 데이터베이스를 탐색 및 목록화하여 주제범위 문헌고찰에 포함된 연구에서 공통으로 제시한 한계점인 (1) '연구에 활용된 데이터 표본검증을 위한 광범위한 데이터의 확보 필요성', (2) '인공지능의 데이터 학습 수준 향상 필요성' 해결을 위한 대안을 제시하고자 하였다.

4.2.2. 후성유전학 관련 분야에서의 인공지능 적용

4.2.2.1. 역학에서의 인공지능 적용현황

역학은 질병의 분포와 결정 요인에 관한 연구로 정의된다(Frérot et al., 2018). 역학 연구는 과학적 방법을 기반으로 하며 데이터 분석에 의존한다. 정의된 모집단에서 다양한 건강 관련 발생을 조사하는 것을 목표로 한다. 데이터 분석은 역학의 가장 중요한 기본 요소이며, 기계학습과 같은 인공지능의 적용을 통

한 계산 능력 향상은 역학 분야의 모델링과 접근 방식을 대폭 확장하였다(Forbes, 2018). 역학에서 인공지능은 질병 유병률 예측 및 분석, 감염병 모니터링, 역학 데이터 처리 및 분석, 실시간 감시 및 응급상황 대응 등에 주로 활용되고 있다.

인공지능 기반 질병 유병률 예측 및 분석 모델은 특히 의료 분야에서 중요한 역할을 할 수 있다. 임상적 진단의 70%는 검사결과에 기반하는데, 인공지능을 활용한 예측모델은 뚜렷한 신체적 증상이 나타나기 전 우려되는 부분을 식별하는 데 도움이 될 수 있으며, 신속한 의료결정을 가능하게 한다. 또한 인공지능을 검사 데이터 워크플로우에 통합함으로써 검사결과를 나이, 성별 등과 같은 환자 정보와 결합하여 질병별 예측모델을 구축할 수 있다. 예를 들어 독일의 지멘스 헬시니어스의 경우, 코로나19에 대응하고자 지난 2021년 환자예측모델인 'Atellica® COVID-19 Severity Algorithm App'을 개발하였다. 딥러닝 기반 모델은 코로나19 임상 중증도 점수를 책정하여 환자의 입원 기간을 예측하고 사망률 등의 통계를 제시하였다(Siemens healthineers., n.d.).

인공지능 기반 역학 감시는 인공지능 기술을 사용하여 전자건강기록, 소셜 미디어 및 뉴스기사와 같은 다양한 출처의 데이터를 확보 및 분석하여 이루어진다(Zeng et al., 2021). 이를 기반으로 신규 혹은 기존 질병에 대한 실시간 모니터링 및 감지, 신규 유행을 정확하게 예측 및 식별하여 질병 유행에 대한 공중보건학적 대응을 가능하게 한다(Anjaria et al., 2023). 미국 질병통제예방센터(Centers for Disease Control and Prevention, CDC)는 'BioSense'라는 기계학습 기반의 시스템을 활용하여 전자의무기록, 응급실 방문, 기타 출처의 데이터를 분석하여 감염병 발생을 식별하였다(Bradley et al., 2005). 캐나다의 'Bluedot'은 의료 전문지식 및 고급 데이터 분석 기술과 인공지능 기술을 이용해 감염병을 추적하고 발생을 예측하였다(Goldust et al., 2023). 'Artificial

Intelligence in Medical Epidemiology(AIME)’ 은 브라질, 싱가포르, 말레이시아에서 주로 활용되는 기계학습 기반 인공지능 모델로서 뎅기열의 유행을 사전에 예측하고 대비 및 완화하기 위한 목적으로 개발되었다. AIME는 뎅기열 유행을 88.7% 정확도로 사전예측하였으며 마닐라, 리우데자네이루를 포함 900만 명의 사람들을 뎅기열 감염으로부터 보호할 수 있었다(Nesta., n.d.). 이외에도 Cognizant, EPIC Systems, eClinicalWorks, Komodo Health, Microsoft 등은 공중보건의 목표 달성을 위해 역학적 측면에서 인공지능 기술을 개발 및 적용해오고 있다(표 4).

표 4. 역학에서의 인공지능 적용현황

번호	회사 및 모델명	인공지능 사용 목적	상세사항	인공지능 기술영역
1	Siemense Healthineer	정보수집, 질병 예측	데이터수집, 임상 중증도 책정, 입원 기간 예측, 사망률 등 통계 제시	추론
2	BioSense	정보수집, 질병 예측	전자의무기록, 기타 출처의 데이터를 분석하여 감염병 발생을 식별	추론
3	Bluedot	감염병 추적 및 질병 예측	COVID-19의 집단 감염 예측, 추적	추론
4	Artificial Intelligence in Medical Epidemiology (AIME)	정보수집, 질병 예측	뎅기열의 유행 예측, 유행 발생 지리적 위치 결정, 질병 발생 데이터 입력 및 수집	추론
5	HealthMap	정보수집 및 분석, 감시	데이터 분석, 질병 유행 모니터링, 바이러스 전파추적	추론
6	Cognizant Evolutionary AI™	예측 및 처방	COVID-19 대응을 위해 데이터 기반 학습 및 예측, 처방적 완화 조치 제안	행동
7	EPIC Systems	임상 업무 보조	생성형 AI를 Electronic Health Record(EHR)에 적용, 의료기록 요약·코딩 제안·환자 치료를 위한 실제 증거 제공	행동
8	eClinicalWorks	정보수집, 문서관리	대화형 Electronic Health Record(EHR). 환자 정보 수집. 문서 식별 및 관리, 임상의 업무 보조	행동

표 4. 역학에서의 인공지능 적용현황(계속)

번호	회사 및 모델명	인공지능 사용 목적	상세사항	인공지능 기술영역
9	Komodo Health	정보수집, 임상보조	데이터 접근성 향상. 데이터 통합·맞춤형 분석 및 생성형 AI 비서	행동
10	Microsoft Cloud for Healthcare	의료정보 수집, 임상 업무 보조	의료데이터 수집, 텍스트 분석, Azure AI Health Bot을 통한 임상관리 및 워크로드 관리·보조	행동
11	MEDITECH Inc.	의료정보 수집, 임상 업무 보조	생성형 AI 활용, 임상 문서 작성 보조, 의사 업무 흐름 개선, 비정형 의료정보 요약 및 의료 공급자 제공	행동

참고:

Azure. (2023). *Microsoft empowers health organizations with generative AI and actionable data analysis.* Retrieved Nov 31, 2023 from https://azure.microsoft.com/en-us/blog/microsoft-empowers-health-organizations-with-generative-ai-and-actionable-data-insights/

Cognizant. (2020). *Using evolutionary AI to deal with COVID-19 and more.* Retrieved Nov 31, 2023 from https://digitally.cognizant.com/using-evolutionary-ai-covid-19-and-more-codex5724/

eClinicalWorks. (2023). *eclinicalWorks Bring ChatGPT and AI Models into EHR and Practice Management Solution.* Retrieved Nov 31, 2023 from https://www.eclinicalworks.com/eclinicalworks-brings-chatgpt-and-ai-models-into-ehr-and-practice-management-solution/

FIERCE. (2023). *Epic taps Microsoft to accelerate generative AI-powered 'copilot' tools to help clinicians save time.* Retrieved Nov 31, 2023 from https://www.fiercehealthcare.com/ai-and-machine-learning/epic-expands-ai-partnership-microsoft-rolls-out-copilot-tools-help

FIERCE. (2023). *Komodo Health unveils new full-stack tool to help clients streamline date analysis.* Retrieved Nov 31, 2023 from https://www.fiercehealthcare.com/health-tech/komodo-health-unveils-new-full-stack-solution-maplab-streamline-data-analytics

MEDITECH. (2023). *MEDITECH announces new AI use cases at customer leadership event.* Retrieved Nov 31, 2023 from https://ehr.meditech.com/news/meditech-announces-new-ai-use-cases-at-customer-leadership-event

Nesta. (n.d.). *Artificial intelligence in Medical Epidemiology.* Retrieved Nov 31, 2023 from https://www.nesta.org.uk/feature/collective-crisis-intelligence-case-studies/artificial-intelligence-medical-epidemiology-aime/

Siemens Healthineers. (n.d.). *Atellica COVID-19 Severity Algorithm App.* Retrieved Nov 31, 2023 from https://atellica-covidalgo.azureedge.net/

Zeng, D., Cao, Z., & Neill, D. B. (2021). Artificial intelligence-enabled public health surveillance—from local detection to global epidemic monitoring and control. In Artificial intelligence in medicine (pp. 437-453). Academic Press.

4.2.2.2. 유전학에서의 인공지능 적용현황

강력한 딥러닝 및 인공지능 알고리즘의 개발로 대규모 기계학습이 가능해지면서 인공지능 기술을 이용한 유전체 빅데이터 분석 연구와 사업이 발전하고 있다. 딥지노믹스(Deep Genomics)는 기계학습에 유전체 빅데이터를 결합하여 정밀의료 서비스 제공을 목표로 인공지능 기술을 개발하고 있다. 특정 질병과 인과관계가 있으나 검출하기 어려운 유전질환 돌연변이 분석을 주력에 두고, 단일 유전자 변이로 인해 유전되는 질환을 집중적으로 다루고 있다(Deep Genomics., n.d.). 특히 수집된 생체·의료데이터의 분석 및 활용, 환자군 분류에 인공지능 모델이 적용되었다. Clark et al(2019)의 연구에서는 심각한 병으로 입원한 영아의 의심되는 유전병 진단을 신속히 도와주는 자동화된 인공지능 모델을 개발하였다. 전자의무기록에서 표현형 데이터를 추출하고 유전병과 관련된 특징을 식별하였으며, 유전병을 유발할 가능성이 있는 병원성 변이를 추출하였다. Langelier et al(2018)의 연구에서는 급성 기도감염의 핵심 요소인 병원체 등으로부터 유래된 염기서열 데이터를 통합·분석하여 급성 기도감염 환자군을 정확히 도출하는 인공지능 모델을 개발하였다.

클리어제네틱스(Clear Genetics)의 'Genetic Information Assistant', Optra Health의 'GeneFAX'와 같이 자동화된 인공지능(환자용 AI 챗봇, 가상 어시스턴트)이 활용되고 있다(Sobia Raza, 2020). 약물개발에서도 아스트라제네카(AstraZeneca), 베네볼란트(Benevolent) AI와 같은 의약품 기업들은 인공지능 기술을 활용하여 유전체학, 화학 및 임상 데이터를 분석하고 잠재적인 약물 표적을 신속하게 발견하고 있다. 글락소스미스클라인(GlaxoSmithKline)은 23andMe 회사 데이터를 인공지능 모델이 학습하게 하여, 약물표적을 개발하고 있다(McCartney, 2018; Ordish et al., 2019). 이외에도 유전학 분야에서 인공지능은 다음과 같이 활용되고 있다(표 5).

표 5. 유전학에서의 인공지능 적용현황

번호	회사 및 모델명	인공지능 사용 목적	상세사항	인공지능 기술영역
1	Ardigen	지식 발견	바이오마커, 미생물 분석 등을 위한 다양한 플랫폼 탐색	인지
2	Benevolent AI	의약품 탐색 및 개발	탐색 및 추론된 생물학적 데이터를 활용한 의약품 탐색	인지
3	Deep Genomics	의약품 탐색	인공지능 기반 의약품 탐색 플랫폼 사용	인지
4	Literome	문헌 탐색	PubMed에서 유전체 지식을 추출하기 위한 자동화된 큐레이션	학습
5	Fabric Genomics	변이 해석	유전체 테스트에 대한 인공지능 기반 해석 플랫폼	학습
6	FDNA	표현형 결정	희귀질병의 얼굴 특징을 분석하고 표현형 중심 변이 우선순위 지정	학습
7	Freenome	암 조기 검출 및 치료	조기암 검출 및 치료를 위한 인공지능 유전체학 적용	추론
8	Google (Brain)	변이 감지	Genomic variants를 감지하기 위한 DeepVariant 도구	추론
9	SOPHiA Genetics	변이 감지	Alamut Genova genome 브라우저를 통해 병원성 예측도구 및 알고리즘 통합	추론
10	Clear Genetics	상담 및 보고	유전자에 관한 환자 상담을 위한 AI 챗봇 사용	행동
11	Perthera	치료결정지원	인공지능 분석을 종합하여 암치료 제안	행동
12	Sequana Health	유전체 편집	디자인된 CRISPR 유전체 편집 시스템	행동

참고:

Ardigen. (n.d.). *Artificial Intelligence and Bioinformatics.* Retrieved Nov 15, 2023 from https://ardigen.com/

Benevolent. (n.d.). *About Us.* Retrieved Nov 15, 2023 from https://www.benevolent.com/

Deep Genomics. (n.d.). *AI Workbench.* Retrieved Nov 15, 2023 from https://www.deepgenomics.com/

Fabric Genomics. (n.d.). *Applications of Fabric Enterprise.* Retrieved Nov 15, 2023 from https://fabricgenomics.com/

Freenome. (n.d.). *Early Cancer Detection.* Retrieved Nov 15, 2023 from https://www.freenome.com/

Poon, H., Quirk, C., DeZiel, C., & Heckerman, D. (2014). Literome: PubMed-scale genomic knowledge base in the cloud. Bioinformatics, 30(19), 2840-2842.

Raza, S. (2020). Artificial intelligence for genomic medicine. Cambridge: PHG Foundation, University of Cambridge.

4.2.2.3. 후성유전학에서의 인공지능 적용현황

앞서 진행한 주제범위 문헌고찰 결과에 따르면, 후성유전학에 인공지능을 적용한 경우는 주로 후성유전학적 발생기전 원인에 따른 질병의 발생 예측, 질병유발 표적 특징 유전자를 가진 환자군의 분류, 후성유전학적 특성을 고려한 치료개입 결과 예측을 위해 인공지능 기술이 적용되고 있었다. 이는 인공지능 기술영역 중 '인지', '학습', '추론'에 해당한다.

이외에도 후성유전학적 요인을 고려한 인공지능 모델의 개발은 꾸준한 노력이 이루어지고 있다. 미국 UCLA 연구팀에서 수행한 최근 연구에서는 암 종양과 관련된 후성유전학적 요인을 바탕으로 인공지능 모델을 개발하였다(Cheng et al., 2023). 개발된 모델은 암 등급이나 병기와 같은 전통적 측정보다 환자 개인의 유전자 발현 패턴을 조사함으로써 여러 암 유형에 걸쳐 환자의 건강 결과를 예측하는 데 활용될 수 있다고 언급하였다. 폭소 테크놀로지스(FOXO Technologies)는 인공지능 기반 후성유전체 바이오마커 연구를 위해 2023년 8월 미국의 DataRobot 회사와 파트너십을 체결하였다. 이를 통해 인공지능의 예측 역량, 높은 데이터 처리량 및 자동화를 적용하여 후성유전 바이오마커를 식별하고 인간의 장수와 관련된 요인을 예측하는 것을 목표로 설정하였다(NS Medical Devices., 2023).

선행연구 및 기타 자료를 통해 후성유전학 관련 학문인 역학·유전학 분야에서 인공지능 기술을 어느 정도 영역까지 적용하고 있는지 확인할 수 있었다. 분야별 인공지능 기술 적용현황을 정리하면 다음과 같다(표 6).

표 6. 역학·유전학·후성유전학에서의 인공지능 기술영역별 적용현황

구분		인공지능의 기술영역			
		인지 (지식습득 및 소통)	학습 (알고리즘 통한 학습·판단)	추론 (미래 사건 예측)	행동 (행위 자동화 및 최적화)
역학		- 질병, 의료 정보 수집 - 관련 문서 수집	- 발생수준 식별 - 데이터 통합, 분류, 작성	- 질병 발생률, 유행 등 예측 - 중증도, 입원 기간 예측 등	- 임상 의사결정 보조 - 챗봇(임상관리)
유전학	유전학	- 바이오마커 플랫폼 탐색 - 의약품 탐색	- 유전체 추출, 결과 분석 - 질병의 표현형 분류	- 유전체 다양성 - 질병의 조기 검출 및 예측 - 병원성 예측 도구	- 챗봇(환자상담) - 질병 치료 방안 제안 - 자동 유전체 편집 시스템
	후성유전학	- 후성유전학적 데이터수집 - 환자 정보 수집	- 발생기전, 병원성 증거 확인 - 환자군 분류	- 환자 건강 결과·치료 결과·생존·사망 예측	- 해당없음

참고:

Ayn de Jesus. (2019). Artificial Intelligence in Epidemiology-Current Use-Cases. Emerj. https://emerj.com/ai-sector-overviews/artificial-intelligence-epidemiology/

Caudai, C., Galizia, A., Geraci, F., Le Pera, L., Morea, V., Salerno, E., ... &Colombo, T. (2021). AI applications in functional genomics. *Computational and Structural Biotechnology Journal* , *19* , 5762-5790.

Cheng, M. W., Mitra, M., & Coller, H. A. (2023). Pan-cancer landscape of epigenetic factor expression predicts tumor outcome. *Communications Biology, 6*(1), 1138.

NS Medical Devices. (2023). *FOXO, DataRobot partner for AI-based epigenetic biomarker research.* Retrieved Nov 31, 2023 from https://www.nsmedicaldevices.com/news/foxo-and-datarobot-partner-for-ai-based-epigenetic-biomarker-research/

Quazi, S. (2022). Artificial intelligence and machine learning in precision and genomic medicine. *Medical Oncology* , *39* (8), 120.

Zeng, D., Cao, Z., & Neill, D. B. (2021). Artificial intelligence-enabled public health surveillance-from local detection to global epidemic monitoring and control. In *Artificial intelligence in medicine* (pp. 437-453). Academic Press.

김지선 기자. (2019.11.19.). [AI시대를 준비한다] 세상을 바꿀 AI혁명의 시작. *전자신문*. https://www.etnews.com/20191119000031

4.2.3. 후성유전 분야에서 활용되고 있는 데이터베이스 및 데이터 세트

후성유전 분야에서는 주로 공개 데이터베이스를 통해 대용량의 데이터 세트를 검색 및 활용한다(Grossman et al., 2016). 후성유전 분야에서 활용되고 있는 유전데이터베이스는 다음과 같이 목록화할 수 있다(표 7).

표 7. 후성유전 분야에서 활용되고 있는 유전데이터베이스 목록

번호	데이터베이스명	설명
1	ENCODE (Encyclopedia of DNA Elements)	- 무료 공개 데이터 소스. ENCODE 및 관련 프로젝트의 모든 실험 메타데이터 및 데이터를 제공 - 원검정 및 후속 분석을 수행하는 데 사용된 재료 및 방법에 대한 기록도 함께 저장
2	NIH ROADMAP Epigenomics Mix Mapping Consortium	- 인간의 세포유형 및 조직에 걸친 히스톤 변형, 염색체 접근성, DNA 메틸화 및 mRNA 발현에 대한 게놈 전체 지도 생성 - ‘Integrative Analysis of 111’ 이라는 제목의 논문과 함께 제공되는 보조 데이터 저장소 역할(Nature, 2015)
3	IHEC(The International Human Epigenome Consortium) Data Portal	- 건강 및 질병과 관련된 포괄적인 참조 후생유전체 세트를 제공. 관련 프로젝트의 공개된 데이터를 검색, 다운로드 가능. McGill Epigenomics Data Coordination Center(EDCC)와 McGill Epigenomics Mapping Center(EMC)에 의해 개발, 유지, 관리 - 우리나라는 국립보건원 질병관리본부 형질연구과에서 아시아 국가 최초로 IHEC에 가입하여 한국인 당뇨/비만 등 만성질환 관련 후성유전체 데이터를 생산하여 IHEC에 공개
4	Deep Data Portal	- 선정된 인간 세포 및 조직의 80개까지 참조 후성유전체를 생성 및 해석 - '국제 인간 후성유전체 컨소시엄'인 IHEC에 혁신적인 후성유전체 연구 데이터를 제공
5	UCSC Genome Browser	- 2000년 7월 7일에 웹사이트에서 공개되었으며, UCSC 게놈 브라우저의 초기 프로토타입과 함께 제공 - 실험적으로 확인된 유전자와 컴퓨터 예측 유전자를 모두 제공하며, 수십 가지의 증거 라인을 함께 제시 - ENCODE 프로젝트의 컨소시엄 기관으로 참여

표 7. 후성유전 분야에서 활용되고 있는 유전데이터베이스 목록(계속)

번호	데이터베이스명	설명
6	WashU Epigenome Browser	- 후성유전학 데이터셋의 시각화, 통합 및 분석을 제공하는 웹 기반 유전자 데이터 탐색 도구 - 사용자가 웹 페이지에서 1D(유전체 특징), 2D(Hi-C 등), 3D(유전체 구조), 4D(시계열) 데이터와 상호작용할 수 있는 새로운 통합 패널 설계가 가능
7	4DGenome	- 염색체(Chromatin)상호작용에 대한 데이터베이스 - 5종 8,038,247개의 크로마틴 상호작용 기록을 포함
8	SGC Epigenetic Chemical Probes	- Structure Genomics Consortium(SGC)라는 글로벌 공공-민간 파트너십에서 관리 - 염색체(Chromatin) 관련하여 후성유전학적 신호 전달에 관여하는 단백질 억제, 화학적 탐색 관련 데이터를 포함
9	EWAS Atlas	- DNA 메틸화에 초점을 맞춘 후성유전 분야 연구의 지식 기반이라고 할 수 있는 데이터베이스 - 112개의 조직/세포를 포함하고 305개의 형질, 1830개의 코호트 및 390개의 온톨로지 개체를 포함하는 329,172개의 고품질 EWAS 연관성을 통합
10	miRBase	- ncRNA에 초점을 맞추어, 공개된 microRNA 시퀀스 및 annotation을 검색할 수 있는 데이터베이스
11	Gene Expression Omnibus(GEO)	- NCBI에서 운영하는 공개 유전체 데이터베이스. Microarray, MGS 등 유전체 데이터 제공 - 3종류의 데이터로 구분: GPL(유전체 데이터의 플랫폼), GSE(Data Series), GSM(Sample)
12	The Cancer Genome Atlas Program(TCGA)	- 20,000개 이상의 원발암을 특징, 33개의 암 유형에 걸친 샘플 포함 - 2006년 생성되었으며, 게놈, 후성유전학, 전사체 및 단백질 데이터를 공개
13	Interferome v2.01	- 2009년 NAR Database Edition에 발표된 데이터베이스 - type I, II and III interferon(IFN) 유전데이터 포함 - 감염성, 염증성 질환 및 암에서 선천적 면역 반응과 같은 질병의 발병에 중요한 유전자 특징의 식별 용이
14	PhenomiR	- 2010년 생성, 질병 및 기타 생물학적 과정에서 차별적으로 조절된 miRNA 발현에 대한 정보 제공 - 542개 연구의 종합 데이터베이스
15	ToppGene Suite	- 무료의 공개 개방 데이터베이스 - 유전자 목록의 기능 풍부화, 기능 주석이나 네트워크 분석을 사용한 후보 유전자 우선순위 부여, 상호작용체 내에서 새로운 후보 유전자의 유사성 및 우선순위를 결정하는 종합 포탈

표 7. 후성유전 분야에서 활용되고 있는 유전데이터베이스 목록(계속)

번호	데이터베이스명	설명
16	ArrayExpress	- 고효율의 기능적 유전체 실험에서 얻어진 데이터를 저장하는 기능적 유전체 데이터 저장소. 연구 집단에서 정보의 재사용을 위해 기능적 유전체 데이터를 제공 - 원시 데이터는 European Nucleotide Archive(ENA)로 전송되며, ENA에서 원시 서열 다운받을 수 있음
17	Infinium HumanMethylation450 BeadChip	- 유전체 전체에 분포한 450,000개 이상의 CpG 부위의 메틸레이션 상태 평가 가능 - 유전체 관련 연구 연합 (GWAS) 연구와 같은 대량의 샘플 집단을 스크리닝하기에 이상적
18	dbEM(database of Epigenetic Modifiers)	- 잠재적인 암 표적으로 간주되는 약 167개의 후성유전학적 연산자/단백질 게놈 정보를 포함 - 수천 개의 종양 샘플, 암 세포주 및 건강한 샘플에서 돌연변이, 유전자 발현과 같은 게놈 정보를 제공
19	EpiFactors database	- 인간과 생쥐 후생유전학 조절자, 복합체 및 다중 세포 유형에서의 발현에 대한 기능적 정보를 제공하기 위해 개발 - EpiFactors는 organoid 형성, 암 유전자 변이 검사 및 Sars-CoV-2 감염 연구와 같은 분야에서 사용
20	GWAS Catalog	- 미국 국립인간유전체연구소(National Human Genome Research Institute, NHGRI)와 유럽 생물정보학연구소(European Bioinformatics Institute, EBI)가 공동작업 - 전장유전체연관성분석(genome-wide association study) 일람표 - 연구저널의 데이터를 기반으로 인간의 다양한 질병 및 형질과 관련된 유전변이 정보를 제공

참고:

Epigenie. (n.d.). *Epigenetic Tools and Databases.* Retrieved Sep 31, 2023 from https://epigenie.com/epigenetic-tools-and-databases/

Marakulina, D., Vorontsov, I. E., Kulakovskiy, I. V., Lennartsson, A., Drabløs, F., & Medvedeva, Y. A. (2023). EpiFactors 2022: Expansion and enhancement of a curated database of human epigenetic factors and complexes. *Nucleic Acids Research, 51*(D1), D564-D570.

Singh Nanda, J., Kumar, R., & Raghava, G. P. (2016). dbEM: A database of epigenetic modifiers curated from cancerous and normal genomes. *Scientific reports, 6*(1), 19340.

이외에도 Chung et al(2022)의 연구를 통해 후성유전 분야에서 인공지능을 활용하여 분석할 수 있는 생활습관, 행동 관련 데이터 세트를 확인해볼 수 있었다. 'UCI-HAR', 'mhealth', 'PAMA2P', 'Opportunity'와 같이 공개적으로 이용 가능한 활동데이터 세트는 센서 데이터 분석을 통해 관성 측정 장치(Inertial Measurement Unit, IMU) 및 목표 인간 활동 인식(Human Activity Recognition, HAR)을 사용하여 측정된 행동 데이터를 포함하고 있었다. 'UCI-HAR'은 UCI 대학교에서 수행한 연구로, 19세에서 48세 사이의 봉사자 30명과 함께 수행되었다. 각 개인은 허리에 스마트폰을 착용하고 6가지 활동(걷기, 계단오르기, 계단내려가기, 일어서기, 앉기, 눕기)을 수행하도록 하여 활동데이터를 수집하였다. 'mHealth' 데이터 수집은 마찬가지로 UCI 대학교에서 수행되었다. 연구 대상자의 가슴, 손목 및 발목에 센서를 부탁하고 기본적인 심장 모니터링, 부정맥 확인, 기본 운동(계단 오르기, 무릎굽히기, 달리기, 점프 등)을 통해 행동 데이터를 수집하였다. 활동, 장소, 환경 등의 개인정보 라이프로그 데이터를 확인할 수 있는 'NICIR-14', 'ImageCLEF'과 같은 데이터 세트를 통해 주로 멀티미디어 데이터의 특성을 가진 시각적 데이터를 확인할 수도 있었다. 하지만 시각적 데이터라는 특성 때문에 이용 가능한 원시 데이터의 수는 양적으로 적으며, 개인정보보호 문제로 인해 수집되는 데이터의 가짓수가 제약되었다. 'ExtraSensory'는 일상생활 속에서 개인 스마트폰과 스마트워치를 통해 데이터를 수집하였다. 60명의 피험자로 구성되었으며, 피험자가 앉고, 눕고, 말하고, 걷고, 섭취하고, 운동하는 행위 등을 약 1분 간격으로 수집하였다. 주로 행동 데이터를 축적하는 데 초점을 맞춘 데이터 세트이다(The ExtraSensory Dataset, n.d.).

그러나 생활습관, 행동 관련 데이터의 경우 유전체 데이터와 같이 연구자들이 비교적 쉽게 접근하여 데이터를 활용할 수 있는 공개적인 대규모 데이터베이스는 확인하기 어려웠다.

4.3. 미래 방향성 제언

앞서 선행연구 및 기타 자료를 중심으로 역학·유전학·후성유전학 각 분야에서 인공지능 기술을 어느 정도 영역까지 적용하고 있는지 확인 및 비교하였다. 또한, 후성유전 분야에서 활용 가능한 데이터베이스와 데이터 세트를 탐색 및 목록화하였다. 이를 바탕으로 향후 후성유전 분야에서 인공지능이 어떻게 적용되고 발전되어야 할 것인지에 대해 제안하고, 선행연구에서 제한점으로 제시된 인공지능 학습데이터의 한계를 극복하기 위한 대안을 다음과 같이 제시하고자 한다.

4.3.1. 데이터 통합성 향상을 위한 결합형 빅데이터 활용 활성화

데이터 통합성 향상 및 표준화를 위해 유전형 데이터와 표현형 데이터를 결합한 빅데이터 도출 및 활용 활성화가 필요하다. 유전자 정보뿐 아니라 병원 기록(질병, 검사 등), 생활습관(식습관, 음주, 흡연, 운동 등), 환경(미세먼지, 오존 등) 등의 데이터를 결합한 분석이 요구된다. 이러한 정보를 인공지능 모델과 통합시킴으로써 변수 간 복잡한 상호작용을 식별하고, 질병 예측, 합병증 예측 역량 향상에 기여할 수 있다(Alshabab et al., 2022).

지금까지 환자에 대한 평가는 인구통계학적 변수, 동반 질환 등에 대한 정보로 제한되어왔다. 이런 요소만으로는 관찰된 건강 결과 중 30%를 설명할 수 없는 경우가 존재한다(Haddad et al., 2023). 기존 연구결과는 특성 유전체와 관련된 발현 메커니즘이 확인되는 집단에 대한 정밀의료개입을 가능하게 하지만 개별 환자의 생활습관과 같은 정보들을 고려하지 못하기 때문에 개인에게 적합한 정밀의료를 제공하는 데에는 한계가 존재한다(Hawkins et al., 2023).

이에 특정 유전자에 초점을 맞추어 질병을 예측하는 것에서 더 나아가, 개인의 생활습관으로 인해 발생 가능한 후성유전학적 변화까지 함께 고려하여 질병을 예측하고 예방할 필요성이 있다. 그러나 앞서 살펴본 바와 같이 실제 후성유전 분야에서 활용 가능한 데이터는 기존에 병원에서 수집된 환자 정보만을 활용하거나, 대규모 코호트를 통해 수집된 유전체 데이터만이 대부분이라 연구자들이 사용할 수 있는 데이터가 한정적이라는 단점이 존재한다. 또한, 연구마다 사용하는 데이터 종류가 다르고 여러 데이터 세트와 데이터베이스들로 구분되어 있어 데이터 간 상호연결성이 떨어진다는 문제점이 있다.

유전자 중심의 검사는 나이가 들면서 변화하지 않고, 질병 위험의 극히 일부만을 통제하는 유전자에 대한 요약만을 제공한다. 반면 후성유전학적 검사는 질병 발생 위험 수준에 영향을 줄 수 있는 변화하는 생활습관에 대한 포괄적인 정보를 제공할 수 있다. 이러한 측면에서 유전형 데이터와 표현형 데이터를 결합한 빅데이터를 수집하는 것이 필요하다. 수많은 데이터가 기기를 통해 생성되고, 전자적으로 저장되어야 하며, 상호연결성 향상을 위해 표준화될 필요성이 있다. 데이터 간의 상관성을 분석하고, 분석결과 기반의 질병 예측 및 위험도 평가 등을 통해 예방적 측면에서 개인에게 적합한 의료개입이나 중재가 가능하도록 하여야 한다 (Haddad et al., 2023).

결합형 빅데이터 세트의 구축 및 활용은 또한 의료기관과 유관 기업들의 데이터 접근성을 확대함으로써 건강관리의 연속성 확보에 기여할 수 있다. 더 나아가 생활 밀착형 질병 관리를 위한 해결책을 제시할 수 있다. 개인 삶의 가치 확대, 생활습관 제시 및 관리, 발병 전 의료개입의 중재 가능성을 확보함으로써 개인 맞춤형 정밀의료를 가능하게 할 수 있다.

4.3.2. 후성유전 분야에서의 인공지능 기술적용 영역 확대

앞서 유전형 데이터와 표현형 데이터를 결합한 빅데이터를 수집 및 활용하는 것이 필요하다고 제안하였다. 이때 유전형 데이터와 표현형 데이터의 수집에 있어 자동화된 인공지능 기술의 적용이 요구된다. 자동화된 인공지능을 통해 인간이 나이가 들면서 환경이 건강에 어떤 영향을 미치는지, 생활습관이 건강에 어떤 영향을 미치는지에 대한 정보를 여러 차례, 여러 시간에 걸쳐 수집할 수 있다 (Gharipour et al., 2021; Lester et al., 2016). 데이터를 단순히 수집하는 것에서 나아가, 인간이 특정 생활습관을 지속하거나 특정 환경에 지속적으로 노출되었을 때, 후성유전학적 측면에서 특정 질병 유전자가 실제로 발현되는지에 대한 정기적인 검증이 요구된다. 이를 위해선 자동화된 인공지능 모델을 통해 주기적으로 개인의 유전형 및 표현형 데이터를 수집 및 분석하고 유전자 발현 여부를 확인할 수 있어야 한다.

후성유전과 관련하여 인공지능을 구현한 선행연구과 동향을 살펴보면, 후성유전 정보를 수집하고, 질병 데이터와 환자 데이터 등을 수집하는 '인지' 기술 측면에서 인공지능이 적용되었다. 또한, 습득한 데이터를 기계학습이나 딥러닝 알고리즘을 활용하여 분석함으로써 '학습' 기술이 활용되고, 질병을 발생시키는 병원성 증거를 확인하고 해당 증거 요인을 가진 환자 데이터를 분류하고, 환자 건강 결과나 치료 결과를 예측함으로써 '추론' 기술영역까지 활용되고 있었다. 그러나 아직 후성유전 분야에서는 추론 결과를 바탕으로 물리적인 행동을 수행하거나 시스템의 처리를 발생시켜 실제적인 작업을 수행하도록 하는 '행동' 기술 측면이 활용되고 있지는 않았다.

역학이나 유전학에서는 인공지능 기술을 활용하여 챗봇을 통한 환자 상담 응대, 질병 및 임상적 데이터를 바탕으로 필요한 치료개입을 판단하여 의사의 임상적

의사결정을 보조하고, 자동화된 유전체 편집시스템을 활용하여 유전적으로 예측 가능한 환자의 질병 발생을 사전 예방하고 건강 결과를 개선하는 등의 작업이 수행되고 있다. 특히 인공지능을 활용한 챗봇, 건강관리 등의 애플리케이션, 스마트폰, 스마트워치와 같은 모바일기기나 인터넷 의료기기(Internet of Medical Things, IoMT)를 통해 사람의 행동 변화를 장려하고 있다(Haiyu et al., 2019; Polu & Polu, 2019). 예를 들어 유전적으로 당뇨 유전자를 가진 환자의 경우 균형감 있는 식습관을 제안하거나, 고혈압 또는 심혈관계 유전자를 가진 환자의 경우 심박 수와 혈압을 주기적으로 측정하여 수치가 높은 경우 경고 문구를 보내는 등의 작업을 수행한다.

이런 측면에서 후성유전 분야에서도 개인에게 적합한 생활습관을 관리하고 예방적 관점의 건강관리가 가능하도록 '행동' 기술영역까지 인공지능의 적용을 확대할 필요성이 있다. 후성유전학적 접근을 통해 기존의 질병 유발 유전자 자체를 바꾸지 못할 순 있어도, 식이나 운동 등과 같은 생활습관에 의해 켜고 꺼지는 유전자를 관리할 수 있다. 예를 들어 비만 유전자를 가진 사람의 경우 인공지능 기반 애플리케이션 혹은 챗봇 등을 통해 건강하고 균형 잡힌 식사 방식을 제안받을 수 있다. 제안받은 식사 방식을 일정 기간 행동한 후, 질병 유발 유전자의 발현 여부를 확인한다. 질병 유발 유전자가 발현되지 않으면 인공지능은 제안하였던 식사 방식과 유사한 방식을 다시금 제안하고, 질병 유발 유전자의 발현 여부를 주기적으로 확인하여 궁극적으로 질병의 발생 자체를 예방하고 모니터링 할 수 있다(Lee et al., 2022).

후성유전이 궁극적인 목표로 하는 정밀의료는 질병의 예방 및 관리까지 통합하는 개념이기에, 후성유전 측면의 자동화된 응용 프로그램은 사용자의 개별 니즈를 중심으로 일상적 건강관리 바탕의 디지털 예방중심 서비스로 작용되어야 할 필요성이 있다.

제5장 고찰 및 결론

5.1. 연구방법에 대한 고찰

본 연구는 주제범위 문헌고찰 방법을 적용하여 후성유전 분야에서 인공지능 기술이 적용된 문헌을 수집하고, 활용되고 있는 인공지능 기술의 종류와 후성유전 분야 관련 데이터 세트를 검토하고자 하였다. 이를 통해 후성유전 분야에서의 인공지능 적용 미래 방향성을 제시하는 것을 목적으로 하였다. 연구결과, 관련 연구가 어떻게 진행되고 있는지, 인공지능이 어느 정도 범위까지 적용되고 있는지 조사할 수 있었다. 그러나 본 연구는 다음과 같은 제한점을 가진다. 첫째, 학술연구정보서비스(RISS), 한국학술정보(KISS), Pubmed를 통한 문헌 검색을 수행하였다. 문헌 포함 수준 확대를 위해 Google Scholar를 통한 수기 검색을 진행하였음에도, 일부 관련 문헌은 확인되지 않았을 가능성이 있다. 둘째, 주제범위 문헌고찰 방법은 광범위한 문헌을 포함하여 포괄적으로 고찰하고자 근거의 질 평가를 거치지 않으므로 잠재적 비뚤림(Bias) 위험이 있을 수 있다.

또한 본 연구는 전통적 문헌고찰 방법을 활용하여 역학·유전학·후성유전학에서 인공지능 기술을 어느 정도 영역까지 적용하고 있는지 탐색하였으며, 비교 결과를 바탕으로 후성유전 분야에서의 인공지능 적용 미래 방향성을 제시할 수 있었다. 그러나 문헌고찰에 포함된 자료의 해석 과정에서 연구자의 주관이 개입되었을 가능성이 있을 수 있다. 이에 본 연구에서 제시한 미래 방향성을 검증하고 주장의 신뢰성 확보, 다양한 실천적 방안 모색을 위해 전문가 심층면담이나 전문가 자문 등의 후속 연구가 요구된다.

5.2. 연구결과에 대한 고찰

본 연구는 후성유전 분야에서 인공지능이 어느 정도 기술영역까지 적용되고 있는지 확인하기 위해 '주제범위 문헌고찰' 방법을 활용하여 현황을 파악하였다. 이와 더불어 전통적 문헌고찰 방법을 활용하여 후성유전 분야에서의 인공지능 적용 미래 방향성을 제시하였다. 미래 방향성으로 제시한 '데이터 통합성 향상을 위한 결합형 빅데이터 활용 활성화', '후성유전 분야에서의 인공지능 기술적용 영역 확대'는 최근 건강관리 관련 개인의 인식 및 관심 증대와 연결된다.

최근에는 예방의학을 넘어 일상적 건강관리가 추세로 자리 잡고 있다. 많은 개인이 디지털 기기를 통해 의료 및 헬스케어에 접근하고자 하고 있으며, 소비자 인식 및 기대 수준이 높아지고 있다. 이러한 측면에서 유전형 데이터 및 표현형 데이터를 결합한 빅데이터 체계 구축은 생활 밀착형 질병 관리 해결책을 제시할 수 있을 것이다.

더불어 우리나라 정부의 '보건의료 빅데이터 플랫폼', '국가 통합 바이오 빅데이터 구축사업'과도 연관된다. '보건의료 빅데이터 플랫폼'이란 보건의료 공공데이터를 결합 및 가명처리 하여 공공 목적의 연구에 활용할 수 있도록 개방하는 시스템이다(김준호, 2023). 이를 통해 질병관리청 국민건강영양조사·KoGES 기반 통합자료·예방접종 데이터베이스·결핵환자신고현황연보 데이터베이스, 통계청 사망원인통계 자료, 국립재활원 의무기록·간호기록 데이터 세트, 건강보험심사평가원 치료내역·상병내역 등, 국립암센터 암등록 데이터 세트, 국립중앙의료원 치매 관련 데이터 세트 등을 제공하고 있다(보건의료 빅데이터 플랫폼, n.d.). 최소 2곳 이상의 제공기관 데이터를 연계·결합하고자 하는 경우 사회적 기여도 등을 입증하는 자료와 함께 데이터 활용신청을 할 수 있다. 활용신청이 접수되면 연구평가위원회·데이터 제공기관 심의 등을 거쳐 보건의료 빅데이터 플랫폼을

통한 연계 데이터를 활용할 수 있다.

그러나 데이터 활용신청 절차나 심의 절차가 복잡하고 연구자가 활용 가능한 데이터가 제한적이라는 한계점을 가진다. 또한 한국인 대상 유전체 정보 중심의 다양한 데이터 결합 필요성을 학계 및 산업계에서 지속적으로 언급해오고 있다(김주연, 2023). 이를 해결하기 위한 수단으로 우리나라는 보건복지부·과학기술정보통신부·산업통상자원부·질병관리청 범부처 사업인 '국가 통합 바이오 빅데이터 구축사업'을 기획, 2023년 6월 예비타당성 조사를 통과하였다(김민준, 2023). '국가 바이오 빅데이터'를 통해 정밀의료 기술개발 등 의료 혁신과 바이오헬스 성장을 달성하는 것을 목표로 한다. 참여자의 자발적 참여와 동의에 기반하여 혈액과 소변 등 검체를 채취하고 유전체, 전사체, 구조데이터 등과 의료활동 및 시험을 통해 산출되는 임상 또는 전임상 정보, 생활(라이프로그) 정보를 포함하는 Omics 데이터를 포함하고자 계획하였다. 여기에 공공 및 개인보유 건강정보를 연계함으로써 개인 중심의 결합형 데이터를 구성하고 관리하고자 하고 있다(이용호 등, 2023).

이런 측면에서, 본 논문에서 미래 방향성으로 제시한 결합형 빅데이터 구축 및 활성화는 우리나라의 보건의료빅데이터 통합 플랫폼 활성화와 국가 바이오 빅데이터 필요성을 뒷받침하는 참고자료로 활용될 수 있을 것이다. 보건의료빅데이터 통합 플랫폼과 국가 바이오 빅데이터에서 제공하는 데이터 간 결합을 추진한다면, 후성유전 분야를 비롯한 유관분야의 연구확대 뿐 아니라 정밀의료의 발전에 기여할 수 있을 것으로 기대된다. 예를 들어 보건의료빅데이터 통합 플랫폼의 '치매 관련 데이터 세트'와 국가 바이오 빅데이터에서 제공하고자 하는 '임상 및 전임상 정보·Omics 데이터·생활(라이프로그) 정보'를 결합한다. 임상, Omics 및 치매데이터 세트를 바탕으로 인공지능을 학습시켜 치매 유전자를 가진 환자군(혹은 예비 환자군)을 분류한다. 분류된 치매 유전자 환자군을 대상으로 특정 기간

의 라이프로그 정보와 치매 유전자 발현 정보를 수집 및 결합한다. 유전자가 발현되었거나 발현되지 않은 경우를 구분하고 데이터를 가공한다. 만일 치매 유전자가 발현된 환자군에서 유사한 특정 생활습관을 행하였거나 유사한 특정 환경에 지속적으로 노출된 사실이 있는 경우, 후성유전학적으로 질병이 발생한 것이라는 결과를 도출해낼 수 있다. 이러한 연구결과를 바탕으로 치매 유전자를 가진 환자에게 발현 조절 또는 억제를 위한 생활습관을 제안할 수 있으며, 환경요인으로부터의 노출을 사전 예방할 수 있을 것이다.

이러한 결합형 빅데이터 세트 마련을 위한 데이터의 수집 및 관리를 위해서는 인공지능 기술을 단순히 적용하는 단계에서 더 나아가, 자동화된 수준의 인공지능 기술이 적용되어야 한다. 이에 개인에게 적합한 생활습관을 관리하고 예방적 관점의 디지털 헬스케어가 가능하도록 인공지능의 적용을 '행동' 기술영역까지 확대할 것이 요구된다.

그러나 인공지능 기술영역 확대에 앞서 유전형 및 표현형 데이터 수집과 활용을 위한 사회·정책적 논의가 필요하다. 유전형 데이터 도출을 위해 필요한 건강 정보에 대한 소유권과 관리 권한은 의료기관이 가지고 있고, 개인의 생활습관이나 환경 데이터와 같은 표현형 데이터는 각 개인으로부터 수집되기에 개인에게 소유권이 존재한다(최경환, 2020). 이에 유전형 데이터 소유권 및 관리 권한의 범위 설정과 명확한 주체를 설정할 필요성이 있다. 표현형 데이터 수집에 있어 개인정보 침해 방지와 표준화된 자료 수집을 위해 수집방법, 장소, 시간, 범위 등 포함하고자 하는 수집 정보의 수준과 범위에 대한 논의가 뒷받침되어야 한다.

이와 함께 후성유전정보의 수집과 활용으로 인해 발생 가능한 차별 문제를 고려하여야 한다. 유전정보로 인한 차별 문제는 인간게놈프로젝트의 시작부터 사회적 부작용으로 거론되었다. 이에 2004년 「생명윤리 및 안전에 관한 법률(이하 '생

명윤리법’)」 제정 당시부터 유전정보를 이유로 보험에서 개인을 차별하는 것을 금지해오고 있다. 그러던 2016년 6월 30일 ‘생명윤리법’을 통해 의료기관이 아닌 유전자검사기관의 유전자검사를 예외적으로 허용하였다(양지현 & 김소윤, 2017; 조수민 등, 2022). 이에 따라 개인의 유전정보에 의한 차별 금지 문제를 온전히 예방하기 어려워졌다. ‘생명윤리법’ 제46조(유전정보에 의한 차별 금지 등)에 ‘유전정보를 이유로 교육・고용・승진・보험 등 사회활동에서 다른 사람을 차별하여서는 아니 된다’고 명시하고 있다. 그러나 광범위하고 원칙적인 규칙만으로는 실제적인 유전정보 활용 영역에서 발생 가능한 문제점을 모두 해결하고 위반행위를 금지하기는 어려우며, 개인의 생체정보를 이용한 건강관리서비스의 연계 관련 개인 및 소비자의 권리 보호에 있어 공백과 한계점이 존재한다.

해외의 경우에도 유전자 정보 기반의 차별 금지를 위한 정책적 기반을 마련한 바 있다. 미국에서는 Genetic Information Non-discrimination Act(이하 ‘GINA’)을 통해 유전적으로 질병에 대한 경향을 가진 사람들에게 건강 보험이나 고용 거부에 대한 보다 강력한 보호를 제공하도록 규제하였으며, 캐나다의 경우 2017년 5월 Genetic Non-Discrimination Act(이하 ‘GNA’)을 제정하여 건강관리자와 연구자를 제외한 어떤 사람이든 개인에게 유전검사를 받거나 유전검사결과를 공개할 것을 요구하지 않도록 규제하였다(Dupras et al., 2018; OpenParliament.ca, 2017). 그러나 GINA와 GNA에서 언급하고 있는 ‘유전정보’의 정의가 ‘DNA’, ‘RNA’, ‘유전자형’이라고 지칭되고 있어 후성유전정보까지 규제할 수 없다는 한계점이 존재한다(Dyke et al., 2019).

2016년 11월, GWG Holdings의 보험 기술 자회사인 Life Epigenetics는 DNA 메틸화 프로파일링을 통해 개인의 기대 수명을 예측할 수 있는 것으로 알려진 후성유전학 기술의 활용에 대한 독점 라이센스를 확보하였다. 이후 2017년 3월 GWG Life 보험회사는 개인의 후성유전학적 정보를 바탕으로 실제 생물학적 연령을 결

정하기 위해 보험 소유자가 제공한 타액 샘플을 수집하기 시작하였다. 이를 통해 일부 보험사가 후성유전학적 정보를 바탕으로 고객에게 보험료나 보장상품의 범위를 차등적으로 설정하는 등 차별이 발생할 수 있다는 우려의 목소리가 대두되었다(Dupras et al., 2018). 이에 최근 미국에서는 GINA에 후성유전학적 정보로 인해 발생가능한 차별에 대한 규제까지 반영되어야 한다는 목소리가 높아지고 있다(Crystal Grant, 2023).

이렇듯 현재 수준의 유전정보 차별금지 관련 정책적 규제는 유전정보 수준 위주로 적용되고 있어, 후성유전 정보 수준에서의 차별 행위를 포함한 문제 행위까지 규제하기 어려운 실정이다. 발생 가능한 여러 한계점을 극복하고 잠재적 문제를 고려하기 위해 윤리적·법적·사회적 문제(Ethical, Legal and Social Issues, 이하 'ELSI')를 고려한 규제 마련이 필요할 것이다.

5.3. 결론

본 연구에서는 후성유전 분야에서의 인공지능 적용과 관련하여 주제범위 문헌고찰을 통해 현황을 파악하고, 전통적 문헌고찰 방법을 활용하여 후성유전 분야에서의 인공지능 미래 적용 방향성을 제시하였다.

최근에는 예방의학을 넘어 일상적인 건강관리가 추세로 자리 잡고 있으며, 많은 개인이 디지털 기기를 통해 의료 및 헬스케어에 접근하고자 하고 있다. 이와 더불어 소비자 인식 및 기대 수준이 높아지고 있다. 그러나 기존의 후성유전 분야 연구결과는 특성 유전체와 관련된 발현 메커니즘이 확인되는 집단에 대한 정밀의료개입을 가능하게 하지만 개별 환자가 가지고 있는 생활습관과 같은 정보들을 고려하지 못하기 때문에 개인에게 아주 적합한 정밀의료를 제공하는 데에는 한계가 존재한다(Hawkins et al., 2023).

이에 본 연구는 기존 연구에서 확인된 한계점을 바탕으로 후성유전 분야에서의 인공지능 적용 미래 방향성을 제시하였다는데 그 의의를 가진다. 본 연구에서 제시한 미래 방향성인 결합형 빅데이터 세트 구축 및 활용을 통해 연구자의 데이터 접근성을 높일 수 있으며, 데이터 표준화를 통한 상호연결성을 확대할 수 있다. 의료기관과 유관 기업들의 데이터 접근성을 확대함으로써 건강관리의 연속성을 확보할 수 있다. 최근 우리나라에서 '보건의료 빅데이터 통합 플랫폼'을 통해 흩어져 있는 여러 데이터들을 연구자가 신청 시 결합하여 제공하거나 분석할 수 있도록 하고 있다는 점, 한국인 유전체 바이오뱅크 마련을 위해 '국가 바이오 빅데이터' 구축을 추진하여 Omics 데이터를 수집 및 관리하고자 계획하고 있다는 점에서 보건의료빅데이터 통합 플랫폼의 활성화 뿐 아니라 국가 바이오 빅데이터의 필요성을 뒷받침하는 참고자료로 활용될 수 있을 것이다.

또한 후성유전 분야에 있어서 인공지능 기술영역 확대의 필요성을 제시하였다. 후성유전 분야에서 자동화된 인공지능 기술영역의 적용은 개인의 생활습관 데이터를 적절한 주기마다 수집하고, 수집된 정보를 자동화된 메커니즘에 의해 분석하여 유전적 요인과의 상호작용을 바탕으로 개인별 건강 행동을 위한 매뉴얼을 제시하도록 할 수 있다. 제시된 매뉴얼을 바탕으로 개인이 건강 행동을 수행하고, 해당 건강 행동이 후성유전학적 발생기전을 유발하는지를 확인, 필요할 경우 매뉴얼을 재생성할 수 있다. 예방적 차원에서 개인의 질병 발생을 낮출 수 있도록 중재 혹은 치료 전 개입이 가능하도록 도움을 줄 수 있을 것이다. 이는 궁극적으로 후성유전 측면에서 개인별 정밀의료의 실현을 가능하게 할 수 있다.

그러나 후성유전 정보 수집 및 활용의 범위를 설정하는 것이 필요하며, 정보 수집으로 인해 발생가능한 차별 문제를 예방하기 위해 ELSI를 고려한 법·윤리적 규제 마련에 대한 사회적 논의가 선행되어야 할 것이다. 이러한 사회적 논의가 선행되었을 때, 개인 맞춤형 정밀의료가 강화될 것이며 궁극적으로 개인 수준의 건강 관리에서 나아가 집단 수준의 건강관리까지 가능하게 하는 공중보건의 목표 달성을 가능하게 할 수 있을 것이다.

참고문헌

A. M. Newman et. al. (2015). Robust Enumeration of Cell Subsets from Tissue Expression Profiles. *Nature Methods, 12*(5), 453.DOI: 10.1038/nmeth.3337

Abrahams, E., Ginsburg, G. S., & Silver, M. (2005). The personalized medicine coalition: goals and strategies. *American Journal of Pharmacogenomics, 5,* 345-355.

Accenture. (2019). *Ai: an engine for growth.* Retrieved Sep 25, 2023 from https://www.accenture.com/us-en/insight-artificial-intelligence-healthcare

Alegría-Torres, J. A., Baccarelli, A., & Bollati, V. (2011). Epigenetics and lifestyle. *Epigenomics, 3*(3), 267-277. https://doi.org/10.2217/epi.11.22

Alshabab, B. S., Lafage, R., Smith, J. S., Kim, H. J., Mundis, G., Klineberg, E., ... & Lafage, V. (2022). Evolution of proximal junctional kyphosis and proximal junctional failure rates over 10 years of enrollment in a prospective multicenter adult spinal deformity database. *Spine, 47*(13), 922-930.

Amazon Web Services. (n.d.). *Radial Kernel SVM.* Retrieved Sep 25, 2023 from https://rstudio-pubs-static.s3.amazonaws.com/170893_9e4e88b0dedc4adfa3bd5a87eb64a9ba.html

analystic steps. (n.d.). *A Classification and Regression Tree(CART) Algorithm.* Retrieved Nov 10, 2023 from https://www.analyticssteps.com/blogs/classification-and-regression-tree-cart-algorithm

Analystics Vidhya. (2023.09.21.). *Adaboost Algorithm: Understand Implement and Master AdaBoost*. Retrieved Sep 25, 2023 from https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners/

Analytics Vidhya. (2023.10.27.). *Guide on Support Vector Machine Algorithm.* Retrieved Sep 25, 2023 from https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/

Anjaria, P., Asediya, V., Bhavsar, P., Pathak, A., Desai, D., & Patil, V. (2023). Artificial Intelligence in Public Health: Revolutionizing Epidemiological Surveillance for Pandemic Preparedness and Equitable Vaccine Access. *Vaccines, 11*(7), 1154.

ArcGIS Enterprise. (n.d.). *Portal for ArcGIS. Generalized Linear Regression.* Retrieved Nov 10, 2023 from https://enterprise.arcgis.com/en/portal/latest/use/geoanalytics-generalized-linear-regression.htm

Ardigen. (n.d.). *Artificial Intelligence and Bioinformatics.* Retrieved Nov 15, 2023 from https://ardigen.com/

Arksey, H., & O'Malley, L. (2005). Scoping studies: towards a methodological framework. *International journal of social research methodology, 8*(1), 19-32.

Arora, I., & Tollefsbol, T. O. (2021). Computational methods and next-generation sequencing approaches to analyze epigenetics data: profiling of methods and applications. *Methods, 187,* 92-103.

Ayn de Jesus. (2019). Artificial Intelligence in Epidemiology-Current Use-Cases. Emerj.

https://emerj.com/ai-sector-overviews/artificial-intelligence-epidemiology/,

Azure. (2023). *Microsoft empowers health organizations with generative AI and actionable data analysis.* Retrieved Nov 31, 2023 from https://azure.microsoft.com/en-us/blog/microsoft-empowers-health-organizations-with-generative-ai-and-actionable-data-insights/

Bae, H. W., Rho, S., Lee, H. S., Lee, N., Hong, S., Seong, G. J., ... & Kim, C. Y. (2014). Hierarchical cluster analysis of progression patterns in open-angle glaucoma patients with medical treatment. *Investigative Ophthalmology & Visual Science, 55*(5), 3231-3236.

Barres, R., Kirchner, H., Rasmussen, M., Yan, J., Kantor, F. R., Krook, A., ... & Zierath, J. R. (2013). Weight loss after gastric bypass surgery in human obesity remodels promoter methylation. *Cell reports, 3*(4), 1020-1027.

Benevolent. (n.d.). *About Us.* Retrieved Nov 15, 2023 from https://www.benevolent.com/

Bird, A. (2002). DNA methylation patterns and epigenetic memory. *Genes & development, 16*(1), 6-21.

Bird, A. (2007). Perceptions of epigenetics. *Nature, 447*(7143).

Bjornsson, H. T., Sigurdsson, M. I., Fallin, M. D., Irizarry, R. A., Aspelund, T., Cui, H., ... & Feinberg, A. P. (2008). Intra-individual change over time in DNA methylation with familial clustering. *Jama, 299*(24), 2877-2883.

Bradley, C. A., Rolka, H., Walker, D., & Loonsk, J. (2005). BioSense: implementation of a national early event detection and situational awareness system. *MMWR Morb Mortal Wkly Rep, 54*(Suppl), 11-19.

Brasil, et al., (2021). Artificial intelligence in epigenetic studies: Shedding light on rare diseases. *Frontiers in Molecular Biosciences,* 8, 648012.

Castro, R., Rivera, I., Struys, E. A., Jansen, E. E., Ravasco, P., Camilo, M. E., ... & Tavares de Almeida, I. (2003). Increased homocysteine and S-adenosylhomocysteine concentrations and DNA hypomethylation in vascular disease. *Clinical chemistry, 49*(8), 1292-1296.

Catboost. (n.d.). *CatBoost is a high-performance open source library for gradient boosting on decision trees* Retrieved Sep 25, 2023 from https://catboost.ai/

Caudai, C., Galizia, A., Geraci, F., Le Pera, L., Morea, V., Salerno, E., ... &Colombo, T. (2021). AI applications in functional genomics. *Computational and Structural Biotechnology Journal , 19* , 5762-5790.

Cavalli, G., & Heard, E. (2019). Advances in epigenetics link genetics to the environment and disease. *Nature, 571*(7766), 489-499. https://doi.org/10.1038/s41586-019-1411-0

Chen, K., & Rajewsky, N. (2007). The evolution of gene regulation by transcription factors and microRNAs. *Nature Reviews Genetics, 8*(2), 93-103.

Cheng, M. W., Mitra, M., & Coller, H. A. (2023). Pan-cancer landscape of epigenetic factor expression predicts tumor outcome. *Communications Biology, 6*(1), 1138.

Chung, S., Jeong, C. Y., Lim, J. M., Lim, J., Noh, K. J., Kim, G., & Jeong, H. (2022). Real-world multimodal lifelog dataset for human behavior study. *ETRI Journal, 44*(3), 426-437.

Clark, M. M., Hildreth, A., Batalov, S., Ding, Y., Chowdhury, S., Watkins, K., ... & Kingsmore, S. F. (2019). Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. *Science translational medicine, 11*(489), eaat6177.

Cognizant. (2020). *Using evolutionary AI to deal with COVID-19 and more.* Retrieved Nov 31, 2023 from https://digitally.cognizant.com/using-evolutionary-ai-covid-19-and-more-codex5724/

Crystal Grant. (2023.05.24.). It's Time for Congress to Update Our Genetic Nondiscrimination Law. *ACLU.* https://www.aclu.org/news/privacy-technology/its-time-for-congress-to-update-our-genetic-nondiscrimination-law

Deafen, D., Escalante, A., Weinrib, L., Horwitz, D., Bachman, B., Roy-Burman, P., ... & Mack, T. M. (1992). A revised estimate of twin concordance in systemic lupus erythematosus. *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology, 35*(3), 311-318.

Deep Genomics. (n.d.). *AI Workbench.* Retrieved Nov 15, 2023 from https://www.deepgenomics.com/

Devereux G, Turner SW, Craig LC, McNeill G, Martindale S, Harbour PJ, et al. Low maternal vitamin E intake during pregnancy is associated with asthma in 5-year-old children. Am J Respir Crit Care Med 2006;174:499-507.

Dupras, C., Song, L., Saulnier, K. M., & Joly, Y. (2018). Epigenetic discrimination: emerging applications of epigenetics pointing to the

limitations of policies against genetic discrimination. Frontiers in Genetics, 9, 202.

Dyke, S. O., Saulnier, K. M., Dupras, C., Webster, A. P., Maschke, K., Rothstein, M., ... & Joly, Y. (2019). Points-to-consider on the return of results in epigenetic research. Genome medicine, 11(1), 1-9.

eClinicalWorks. (2023). *eclinicalWorks Bring ChatGPT and AI Models into EHR and Practice Management Solution.* Retrieved Nov 31, 2023 from https://www.eclinicalworks.com/eclinicalworks-brings-chatgpt-and-ai-models-into-ehr-and-practice-management-solution/

elastic. (n.d.). *자연어 처리(NLP)란 무엇인가?.* Retrieved Nov 31, 2023 from https://www.elastic.co/kr/what-is/natural-language-processing

Enguehard, J., O'Halloran, P., & Gholipour, A. (2019). Semi-supervised learning with deep embedded clustering for image classification and segmentation. *Ieee Access, 7*, 11093-11104.

Epigenie. (n.d.). *Epigenetic Tools and Databases.* Retrieved Sep 31, 2023 from https://epigenie.com/epigenetic-tools-and-databases/

Eraslan, G., Avsec, Ž., Gagneur, J., & Theis, F. J. (2019). Deep learning: new computational modelling techniques for genomics. *Nature Reviews Genetics, 20*(7), 389-403.

Ertel, W. (2019). Artificial Intelligence and Society.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., ... & Dean, J. (2019). A guide to deep learning in healthcare. *Nature medicine, 25*(1), 24-29.

Fabric Genomics. (n.d.). *Applications of Fabric Enterprise.* Retrieved Nov 15, 2023 from https://fabricgenomics.com/

Fardi, M., Solali, S., & Hagh, M. F. (2018). Epigenetic mechanisms as a new

approach in cancer treatment: An updated review. *Genes & diseases, 5*(4), 304-311.

Feinberg, A. P. (2018). The key role of epigenetics in human disease prevention and mitigation. *New England Journal of Medicine, 378*(14), 1323-1334.

FIERCE. (2023). *Epic taps Microsoft to accelerate generative AI-powered 'copilot' tools to help clinicians save time.* Retrieved Nov 31, 2023 from https://www.fiercehealthcare.com/ai-and-machine-learning/epic-expands-ai-partnership-microsoft-rolls-out-copilot-tools-help

FIERCE. (2023). *Komodo Health unveils new full-stack tool to help clients streamline date analysis.* Retrieved Nov 31, 2023 from https://www.fiercehealthcare.com/health-tech/komodo-health-unveils-new-full-stack-solution-maplab-streamline-data-analytics

Forbes. (2018). The rise in computing power: why ubiquitous artificial intelligence is now a reality. *Forbes,* July 17. https://www.forbes.com/sites/intelai/2018/07/17/the-rise-in-computing-power-why-ubiquitous-artificial-intelligence-is-now-a-reality/#22a73011d3f3

Fraga, M. F., & Esteller, M. (2002). DNA methylation: a profile of methods and applications. *Biotechniques, 33*(3), 632-649.

Freenome. (n.d.). *Early Cancer Detection.* Retrieved Nov 15, 2023 from https://www.freenome.com/

Frérot, M., Lefebvre, A., Aho, S., Callier, P., Astruc, K., & Aho Glélé, L. S. (2018). What is epidemiology? Changing definitions of epidemiology 1978-2017. *PloS one, 13*(12), e0208442.

Gharipour, M., Mani, A., Amini Baghbahadorani, M., de Souza Cardoso, C. K., Jahanfar, S., Sarrafzadegan, N., ... & Silveira, E. A. (2021). How are epigenetic modifications related to cardiovascular disease in older adults?. *International Journal of Molecular Sciences, 22*(18), 9949.

Gilmour, D. S., & Lis, J. T. (1985). In vivo interactions of RNA polymerase II with genes of Drosophila melanogaster. M*olecular and cellular biology, 5*(8), 2009-2018.

Goldust, Y., Sameem, F., Mearaj, S., Gupta, A., Patil, A., & Goldust, M. (2023). COVID-19 and artificial intelligence: Experts and dermatologists perspective. *Journal of cosmetic dermatology, 22*(1), 11-15.

Guan, Z., Giustetto, M., Lomvardas, S., Kim, J. H., Miniaci, M. C., Schwartz, J. H., ... & Kandel, E. R. (2002). Integration of long-term-memory-related synaptic plasticity involves bidirectional regulation of gene expression and chromatin structure. *Cell, 111*(4), 483-493.

Guidance, W. H. O. (2021). Ethics and governance of artificial intelligence for health. *World Health Organization.*

Habuza, T., Navaz, A. N., Hashim, F., Alnajjar, F., Zaki, N., Serhani, M. A., & Statsenko, Y. (2021). AI applications in robotics, diagnostic image analysis and precision medicine: current limitations, future trends, guidelines on CAD systems for medicine. *Informatics in Medicine Unlocked, 24*, 100596.

Haddad, S., Pizones, J., Raganato, R., Safaee, M. M., Scheer, J. K., Pellisé, F., & Ames, C. P. (2023). Future Data Points to Implement in Adult Spinal Deformity Assessment for Artificial Intelligence Modeling Prediction: The Importance of the Biological Dimension. *International*

journal of spine surgery, 17(S1), S34-S44.

Hamamoto, R., Komatsu, M., Takasawa, K., Asada, K., & Kaneko, S. (2019). Epigenetics analysis and integrated analysis of multiomics data, including epigenetic data, using artificial intelligence in the era of precision medicine. *Biomolecules, 10*(1), 62.

Haoyu, L., Jianxing, L., Arunkumar, N., Hussein, A. F., & Jaber, M. M. (2019). An IoMT cloud-based real time sleep apnea detection scheme by using the SpO2 estimation supported by heart rate variability. *Future Generation Computer Systems, 98*, 69-77.

HapMap Consortium, T. I. (2003). The international HapMap project. *Nature, 426*(6968), 789-796.

Hawkins-Hooker, A., Visonà, G., Narendra, T., Rojas-Carulla, M., Schölkopf, B., & Schweikert, G. (2023). Getting personal with epigenetics: towards individual-specific epigenomic imputation with machine learning. *Nature Communications, 14*(1), 4750.

Holder, L. B., Haque, M. M., & Skinner, M. K. (2017). Machine learning for epigenetics and future medical applications. *Epigenetics, 12*(7), 505-51.

IBM. (2022). *C5.0 노드.* Retrieved Sep 25, 2023 from https://www.ibm.com/docs/ko/

IBM. (n.d.). *What is the k-nearest neighbors algorithm?.* Retrieved Sep 25, 2023 from https://www.ibm.com/docs/ko/

IBM. (n.d.). *랜덤 포레스트란?.* Retrieved Sep 25, 2023 from https://www.ibm.com/kr-ko/topics/random-forest

Insights, F. (2019). AI and healthcare: a giant opportunity. *Forbes.* February, 11.

Jaenisch, R., & Bird, A. (2003). Epigenetic regulation of gene expression:

how the genome integrates intrinsic and environmental signals. *Nature genetics, 33*(3), 245-254.

Jaffe, A. E., Murakami, P., Lee, H., Leek, J. T., Fallin, M. D., Feinberg, A. P., & Irizarry, R. A. (2012). Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies. *International journal of epidemiology, 41*(1), 200-209.

JĘdrychowski, W., GaŁŚ, A., Whyatt, R., & Perera, F. (2006). The prenatal use of antibiotics and the development of allergic disease in one year old infants. A preliminary study. *International Journal of Occupational Medicine & Environmental Health, 19*(1).

Jenuwein, T. (2006). The epigenetic magic of histone lysine methylation: delivered on 6 July 2005 at the 30th FEBS Congress in Budapest, Hungary. *The FEBS journal, 273*(14), 3121-3135.

Jin, B., Li, Y., & Robertson, K. D. (2011). DNA methylation: superior or subordinate in the epigenetic hierarchy?. *Genes & cancer, 2*(6), 607-617.

Jung, J. Y., & Park, D. (2022). Are AI models explainable, interpretable, and understandable?. *In Human-Centered Artificial Intelligence* (pp. 3-16). Academic Press.

Kaplan, A., & Haenlein, M. (2019). Siri, Siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implications of artificial intelligence. *Business horizons, 62*(1), 15-25.

Kim, H. J., Bang, M., Park, C. I., & Lee, S. H. (2023). Altered DNA Methylation of the Serotonin Transporter Gene Associated with Early Life Stress and White Matter Microalterations in Korean Patients with Panic Disorder. *Neuropsychobiology.*

Kit, A. H., Nielsen, H. M., & Tost, J. (2012). DNA methylation based biomarkers: practical considerations and applications. *Biochimie, 94*(11), 2314-2337.

Kouzarides, T. (2007). Chromatin modifications and their function. *Cell, 128*(4), 693-705.

Krittanawong, C., Zhang, H., Wang, Z., Aydar, M., & Kitai, T. (2017). Artificial intelligence in precision cardiovascular medicine. *Journal of the American College of Cardiology, 69*(21), 2657-2664.

Kumar, H., Lund, R., Laiho, A., Lundelin, K., Ley, R. E., Isolauri, E., & Salminen, S. (2014). Gut microbiota as an epigenetic regulator: pilot study based on whole-genome methylation analysis. *MBio, 5*(6), 10-1128.

Langelier, C., Kalantar, K. L., Moazed, F., Wilson, M. R., Crawford, E. D., Deiss, T., ... & DeRisi, J. L. (2018). Integrating host response and unbiased microbe detection for lower respiratory tract infection diagnosis in critically ill adults. *Proceedings of the National Academy of Sciences, 115*(52), E12353-E12362.

Lappalainen, T., & Greally, J. M. (2017). Associating cellular epigenetic models with human phenotypes. *Nature Reviews Genetics, 18*(7), 441-451.

Lawless, M. W., O'Byrne, K. J., & Gray, S. G. (2009). Oxidative stress induced lung cancer and COPD: opportunities for epigenetic therapy. *Journal of cellular and molecular medicine, 13*(9a), 2800-2821.

Lee, M., Yoon, S. D., Shin, J., Kim, J., & Lee, S. H. (2022). Development of AI-based healthcare system of precision nutrition for health (PNH).

Lester, B. M., Conradt, E., & Marsit, C. (2016). Introduction to the special section on epigenetics. *Child Development, 87*(1), 29-37.

LightGBM. (n.d.). *Read the Docs. Welcome to LightGBM's documentation!.*

Retrieved Sep 25, 2023 from https://lightgbm.readthedocs.io/en/stable/

Lomba, A., Milagro, F. I., García-Díaz, D. F., Marti, A., Campión, J., & Martínez, J. A. (2010). Obesity induced by a pair-fed high fat sucrose diet: methylation and expression pattern of genes related to energy homeostasis. *Lipids in health and disease, 9*(1), 1-10.

Marakulina, D., Vorontsov, I. E., Kulakovskiy, I. V., Lennartsson, A., Drabløs, F., & Medvedeva, Y. A. (2023). EpiFactors 2022: Expansion and enhancement of a curated database of human epigenetic factors and complexes. *Nucleic Acids Research, 51*(D1), D564-D570.

McCartney, M. (2018). Margaret McCartney: AI in medicine must be rigorously tested. *Bmj,* 361.

McGee, S. L., Fairlie, E., Garnham, A. P., & Hargreaves, M. (2009). Exercise-induced histone modifications in human skeletal muscle. *The Journal of physiology, 587*(24), 5951-5958.

McKinsey & Company. (2023.04.24.). *What is AI(Artificial Intelligence)?.* Retrieved Nov 25, 2023 from https://www.mckinsey.com/featured-insights/mckinsey-explainers/what-is-ai

MEDITECH. (2023). *MEDITECH announces new AI use cases at customer leadership event.* Retrieved Nov 31, 2023 from https://ehr.meditech.com/news/meditech-announces-new-ai-use-cases-at-customer-leadership-event

Miller, R. L., & Ho, S. M. (2008). Environmental epigenetics and asthma: current concepts and call for studies. *American journal of respiratory and critical care medicine, 177*(6), 567-573.

Mitchell, T. M. (1997). Machine learning.

Nanney, D. L. (1958). Epigenetic control systems. *Proceedings of the National Academy of Sciences, 44*(7), 712-717.

Neftci, E. O., & Averbeck, B. B. (2019). Reinforcement learning in artificial and biological systems. *Nature Machine Intelligence, 1*(3), 133-143.

Nesta. (n.d.). *Artificial intelligence in Medical Epidemiology.* Retrieved Nov 31, 2023 from https://www.nesta.org.uk/feature/collective-crisis-intelligence-case-studies/artificial-intelligence-medical-epidemiology-aime/

Nesta. (n.d.). *Artificial intelligence in Medical Epidemiology*. Retrieved Nov 31, 2023 from https://www.nesta.org.uk/feature/collective-crisis-intelligence-case-studies/artificial-intelligence-medical-epidemiology-aime/

Nicoglou, A., & Merlin, F. (2017). Epigenetics: A way to bridge the gap between biological fields. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 66,* 73-82.

NS Medical Devices. (2023). *FOXO, DataRobot partner for AI-based epigenetic biomarker research.* Retrieved Nov 31, 2023 from https://www.nsmedicaldevices.com/news/foxo-and-datarobot-partner-for-ai-based-epigenetic-biomarker-research/

Ordish, J., Hannah, M., & Allison, H. (2019). Algorithms as medical devices. *PHG Foundation.*

Polu, S. K., & Polu, S. K. (2019). IoMT based smart health care monitoring system. *International Journal for Innovative Research in Science & Technology, 5*(11), 58-64.

Poon, H., Quirk, C., DeZiel, C., & Heckerman, D. (2014). Literome: PubMed-scale genomic knowledge base in the cloud. *Bioinformatics, 30*(19), 2840-2842.

Qlik. (n.d.). *What is AI Analytics? How It Works & Examples.* Retrieved Nov 31, 2023 from https://www.qlik.com/us/augmented-analytics/ai-analytics

Quazi, S. (2022). Artificial intelligence and machine learning in precision and genomic medicine. *Medical Oncology* , *39* (8), 120.

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine, 380*(14), 1347-1358.

Rasmussen, C. E., & Williams, C. K. (2006). *Gaussian processes for machine learning* (Vol. 1, p. 159). Cambridge, MA: MIT press.

Rauschert, S., Raubenheimer, K., Melton, P. E., & Huang, R. C. (2020). Machine learning and clinical epigenetics: a review of challenges for diagnosis and classification. *Clinical epigenetics, 12*(1), 1-11.

Raza, S. (2020). Artificial intelligence for genomic medicine. *Cambridge: PHG Foundation, University of Cambridge.*

rdrr.io. (2022.12.28.). *stepwiseCOX: Stepwise Cox Proportional Hazards Regression.* Retrieved Sep 25, 2023 from https://rdrr.io/cran/StepReg/man/stepwiseCox.html

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence, 1*(5), 206-215.

Russo, V. E., Martienssen, R. A., & Riggs, A. D. (1996). Epigenetic mechanisms of gene regulation. *Cold Spring Harbor Laboratory Press.*

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of research and development, 3*(3), 210-229.

Sayed, D., & Abdellatif, M. (2011). MicroRNAs in development and disease. *Physiological reviews, 91*(3), 827-887.

ScienceDirect. (2007). *Logistic Regression.* Retrieved Sep 25, 2023 from *https://www.sciencedirect.com/topics/computer-science/logistic-regression*

Sharma, M., Savage, C., Nair, M., Larsson, I., Svedberg, P., & Nygren, J. M. (2022). Artificial intelligence applications in health care practice: scoping review. *Journal of medical Internet research, 24*(10), e40238.

Siemens Healthineers. (n.d.). *Atellica COVID-19 Severity Algorithm App.* Retrieved Nov 31, 2023 from https://atellica-covidalgo.azureedge.net/

Silman, A. J., MacGregor, A. J., Thomson, W., Holligan, S., Carthy, D., Farhan, A., & Ollier, W. E. R. (1993). Twin concordance rates for rheumatoid arthritis: results from a nationwide study. *Rheumatology, 32*(10), 903-907.

Singh Nanda, J., Kumar, R., & Raghava, G. P. (2016). dbEM: A database of epigenetic modifiers curated from cancerous and normal genomes. *Scientific reports, 6*(1), 19340.

Stack Exchange. (2015). *KNN: 1-nearest neighbor*. Retrieved Sep 23, 2023 from https://stats.stackexchange.com/questions/151756/knn-1-nearest-neighbor

Talukder, A., Barham, C., Li, X., & Hu, H. (2021). Interpretation of deep learning in genomics and epigenomics. *Brief Bioinform, 22*(3). https://doi.org/10.1093/bib/bbaa177

Tang, W. Y., & Ho, S. M. (2007). Epigenetic reprogramming and imprinting in origins of disease. *Reviews in Endocrine and Metabolic Disorders, 8,* 173-182.

Tarca, A. L., Carey, V. J., Chen, X. W., Romero, R., & Drăghici, S. (2007). Machine learning and its applications to biology. *PLoS computational biology, 3*(6), e116.

TIBCO. (n.d.). *Stepwise Model Builder-Cox Regression Introductory Overview.* Retrieved Sep 25, 2023 from https://docs.tibco.com/pub/stat/14.0.1/doc/html/UsersGuide/user-guide/stepwise-model-builder-cox-regression-introductory-overview.htm

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., ... & Kalush, F. (2001). The sequence of the human genome. *science, 291*(5507), 1304-1351.

Waddington, C. H. (1942). The epigenotype. *Endeavour, 1,* 18-20.

Woodson, K., Mason, J., Choi, S. W., Hartman, T., Tangrea, J., Virtamo, J., ... & Albanes, D. (2001). Hypomethylation of p53 in peripheral blood DNA is associated with the development of lung cancer. *Cancer Epidemiology Biomarkers & Prevention, 10*(1), 69-74.

xgboost. (2022). *read the docs. Introduction to Boosted Trees-xgboost 2.0.0.* Retrieved Sep 25, 2023 from https://xgboost.readthedocs.io/en/stable/

xilinx. (n.d.). Deep Learning Training vs. Inference: What's the Difference?. https://www.xilinx.com/applications/ai-inference/difference-between-deep-learning-training-and-inference.html

Yang, I. V., & Schwartz, D. A. (2011). Epigenetic control of gene expression in the lung. *American journal of respiratory and critical care medicine, 183*(10), 1295-1301.

Yu, K. H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in

healthcare. *Nature biomedical engineering, 2*(10), 719-731.

Zeng, D., Cao, Z., & Neill, D. B. (2021). Artificial intelligence-enabled public health surveillance-from local detection to global epidemic monitoring and control. In *Artificial intelligence in medicine* (pp. 437-453). Academic Press.

Zhang, F. F., Cardarelli, R., Carroll, J., Zhang, S., Fulda, K. G., Gonzalez, K., ... & Santella, R. M. (2011). Physical activity and global genomic DNA methylation in a cancer-free population. *Epigenetics, 6*(3), 293-299.

강길전. (2010). 후성유전학과 에너지의학. *한국정신과학회 학술대회논문집,* 49-60.

공공데이터의 제공 및 이용활성화에 관한 법률 제1장 제1조.

김민준 기자. (2023.06.30.). '국가 통합 바이오 빅데이터 구축사업' 예타 통과...본격 추진. *메디포뉴스.* https://www.medifonews.com/mobile/article.html?no=180225

김주연 기자. (2023.07.19.). 정부 빅데이터 개방 늘리지만 현장은 '시큰둥' "필요한 자료여야". *청년의사.* http://www.docdocdoc.co.kr/news/articleView.html?idxno=3007776

김지선 기자. (2019.11.19.). [AI시대를 준비한다] 세상을 바꿀 AI혁명의 시작. *전자신문.* https://www.etnews.com/20191119000031

김휘영. (2018). 후생유전학의 기계학습과 의료 응용. *KOSEN 21.* https://kosen.kr/info/kosen/REPORT_0000000000810

류제운, 이상철, 유재수, 김학용. (2013). 노화 관련 유전자의 후성유전학적 특성 분석. *한국콘텐츠학회논문지, 13*(8), 466-473.

문상선. (2023.06.29.). *Image Segmentation 이란?-정의, 종류, 응용분야, 딥러닝, 트렌드.* Datahunt. Retrieved Sep 25, 2023 from

https://www.thedatahunt.com/trend-insight/image-segmentation

박성우. (2018). 우울증의 후성유전기전: BDNF 유전자의 히스톤 변형 및 DNA 메틸화의 역할. *생명과학회지, 28*(12), 1536-1544.

박소연, 이창엽, & 안찬형. (2020). AI, 현재와 미래 - 1부. 인공지능 기술은 어떻게 분류되는가?. *투이컨설팅.* https://www.2e.co.kr/news/articleView.html?idxno=300957.

박일호, & 이흥만. (2018). 비용의 발생에 대한 후성유전학적 관점. *Journal of Rhinology, 25*(1), 1-6.

배다정, & 박춘식. (2013). 천식과 후성유전학. *Allergy, Asthma & Respiratory Disease, 1*(1), 4-10.

보건의료 빅데이터 통합 플랫폼. (n.d.). *데이터 카탈로그 소개.* Retrieved Nov 25, 2023 from https://hcdl.mohw.go.kr/static/cat/catalogIntro

송은모, 조홍석, 김고운, 조재흥, 박히준, & 송미연. (2020). 만성 통증과 후성유전학에 대한 문헌 고찰. *Journal of Korean Medicine, 30*(1).

오정환, 권용대, 윤병욱, & 최병준. (2008). 후생유전학(Epigenetics)과 DNA methylation 의 이해. *대한악안면성형재건외과학회지, 30*(3), 302-309.

이명희. (2002). *혼합 모형을 이용한 microarray 자료의 특이 발현변이 유전자의 추정방법에 관한 연구* (Doctoral dissertation, 연세대학교 대학원).

이용호, 이준학, 강효진. (2023.06.28.). 국가 바이오 빅데이터 인프라의 미래: 바이오 빅데이터 인프라 구축 동향 및 발전방향. *KISTI Issue Brief. 제58호.*

이주하, 김해림, 이상헌, & 김호연. (2013). 류마티스질환에서 후성유전체 변화. *Journal of Rheumatic Diseases (구 대한류마티스학회지), 20*(3), 140-148.

이한상, 박민석, 김준모. (2014). 의료영상에서의 딥 러닝. *대한의학영상정보학회지 2014; 20:*13-18

장안수. (2013). 후성유전과 알레르기 질환. *대한내과학회지, 85*(3), 260-266.

정보영. (2018). *A Study on the Improvement for big data utilization-Focused*

on health insurance claims data (Doctoral dissertation, 연세대학교).

조민호. (2021). 인공지능의 역사, 분류 그리고 발전 방향에 관한 연구. *한국전자통신학회 논문지, 16*(2), 307-312.

천희란, 김수현, & 박은자. (2022). 우리나라 헬스리터러시 측정 도구의 연구 동향 분석: 주제범위 문헌고찰 (Scoping review). *보건교육건강증진학회지, 39*(4), 39-53.

최경환 기자. (2020.05.). 환자만을 위한 헬스케어는 옛말 '개인 유전체 분석' 등 신시장 후끈. *동아비즈니스리뷰.* https://dbr.donga.com/article/view/1206/article_no/9593/ac/magazine

최창현 기자. (2020.01.02.).자연언어처리(NLP) 무엇인가. 그 기술과 시장은?. *인공지능신문.* https://www.aitimes.kr/news/articleView.html?idxno=15036

한국바이오협회. (2020). 바이오 빅데이터(Bio BigData) - 데이터가 생명을 살린다(Data Saves Lives). *Bio economy report.*

한남식. (2015.02.11.). 후성유전체에 대한 연구동향. *BRIC View 동향리포트.* https://www.ibric.org/bric/trend/bio-report.do?mode=view&articleNo=8691676&title=%ED%9B%84%EC%84%B1%EC%9C%A0%EC%A0%84%EC%B2%B4%EC%97%90+%EB%8C%80%ED%95%9C+%EC%97%B0%EA%B5%AC+%EB%8F%99%ED%96%A5#!/list

부 록

<부록 1> 주제범위 문헌고찰 결과 활용되고 있는 인공지능 알고리즘

번호	알고리즘	정의 및 특성	관련 연구
1	Linear-kernel support vector machines(커널 서포트 벡터 머신)	- 두 범주를 분류하면서 마진(margin)이 최대화된 초평면(hyper plane)을 찾는 기법 - 원 공간(Input Space)의 데이터를 고차원공간으로 매핑한 뒤 범주 분류	Hess et al (2020)
2	radial-kernel support vector machines(방사형 커널 서포트 벡터머신)	- 데이터의 선형적 분리 불가 시 좋은 접근법 - 특성에 대한 비선형 변환을 수행하고 고차원공간으로 변환하여, 비선형 데이터를 분리 - 최상의 값을 찾기 위해 교차 검증을 구현하여 조정 매개 변수를 측정할 수 있음	
3	K-Nearest Neighbors, KNN; 1(One)-Nearest Neighbors, 1NN(OneNN) (K-최근접 이웃)	- 비모수적이고 지도된 학습 분류기로, 근접성을 사용하여 개별 데이터 포인트의 그룹화에 대한 분류 또는 예측을 수행 - 유사한 데이터 포인트가 서로 가까이 있다는 가정 바탕. 분류 레이블은 다수결에 근거하여 할당. 즉, 주어진 데이터 포인트 주위에 가장 자주/많이 표시되는 레이블이 함께 분류	Chakravarthy et al (2016); Tran et al (2022)
4	generalized linear model (일반화 선형모형)	- 설명변수 집합과의 관계에 따라 예측을 생성하거나 종속변수를 모델링. 종속변수가 정규분포하지 않는 경우를 포함하는 선형모형의 확장 - glm()함수 사용. family라는 인수를 지정하여 family에 따라 연결된 함수가 달라짐	Shokhirev & Johnson (2022)
5	LightGBM	- Gradient Boosting 프레임워크로 의사결정나무 기반 학습 알고리즘 - 속도가 빠르며 결과의 정확도에 초점을 맞추며 GPU 학습을 지원	Kalyakulina et al (2022)
6	CatBoost	- 의사결정나무 Gradient Boosting을 위한 고성능 오픈소스 라이브러리 - 데이터를 사전 처리하거나 변환할 필요 없는 범주형 기능 지원, 대규모 데이터셋의 경우 다중 구성을 사용하여 빠르고 확장성이 뛰어남	

번호	알고리즘	정의 및 특성	관련 연구
7	Logistic Regression, LR (로지스틱 회귀분석)	- 입력변수가 주어진 이산 결과의 확률을 모형화하는 과정. 가장 일반적인 로지스틱 회귀분석 - 새 표본이 범주에 가장 적합한지 여부를 확인하는 분류 문제에 유용한 분석 방법	
8	AdaBoost	- 이진분류에 사용되는 기술. 여러 개의 약한 학습자를 강한 학습자로 변환하여 예측 능력 향상 - 훈련 데이터셋에 모델을 구축, 첫 번째 모델에서 발생한 오류를 수정하기 위해 두 번째 모델을 구축. 여러 모델을 결합하여 최종 출력을 얻는 방식으로 작동. 각 인스턴스에 대해 가중치를 다시 할당하여 잘못 분류된 인스턴스에 더 높은 가중치를 부여하는 방식으로 작동	Bendifallah et al (2022)
9	eXtreme Gradient Boosting (XGBoost)	- 지도학습 문제에 사용, 훈련 데이터를 사용하여 목표 변수를 예측. 분산환경에서도 실행할 수 있도록 구현한 라이브러리 - 여러 개의 의사결정 나무를 조합하여 사용하는 앙상블 알고리즘: CART 세트	Bendifallah et al(2022); Kalyakulina et al(2022)
10	Elastic net-Cox proportional hazards	- elastic net 페널티로 정규화된 Cox 회귀모델 - 일련의 특징을 최적화하기 위해 매개변수의 최적값을 식별하는 데 적용	Huan et al (2019)
11	Random survival forest (랜덤 생존 포레스트)	- 생존분석을 위한 랜덤 포레스트 방법을 기반으로 하는 앙상블 트리 모델 - 교차 검증을 통해 테스트 된 랜덤 포레스트 모델의 정확도 최대화를 위해 활용	
12	Cox-nnet	- 생존분석을 위한 인공 신경망 기반 방법 - 특징선택을 위해 특징 중요도 점수를 계산	
13	DeepSurv	- 심층 학습 기반 생존 예측 방법 - 다층 피드 포워드 신경망 사용, 숨겨진 레이어는 노드가 포함된 완전 연결 레이어로 이루어진 모델	
14	Lasso Regression (라쏘 회귀)	- 선형 회귀의 또 따른 규제된 버전 - 덜 중요한 특성의 가중치를 제거	Cheng et al(2023); Arabyarmohammadi et al (2022)

번호	알고리즘	정의 및 특성	관련 연구
15	stepwise COX Regression	- 사용자 선택 예측 변수기반 모델식별 - 예측에 대한 통계적 유의성 기준과 정책 및 기타 기준을 사용하여 가장 중요한 예측 변수를 한 번에 한 단계씩 수동으로 회귀방정식하여 모형구축	Cheng et al(2023)
16	Classification and Regression Tree(CART)	- Gini index를 기반으로 의사결정나무를 구축하기 위해 필요한 분류 알고리즘의 한 종류 - 분류트리(트리를 사용하여 대상 변수가 분류에 속할 가능성이 가장 높은 동급을 찾음)와 회귀트리(연속형 변수의 값 예측)를 나타내는 포괄적 단어	de Gonzalo-Calvo et al(2020)
17	Prediction Analysis for Microarrays, PAM	- 유전 발현 데이터를 사용한 클래스 예측을 위한 통계기술. 가까운 중심을 사용하여 클래스를 표현하는 유전자 하위 집합을 식별 - 분류에 활용할 수 있으며, 생존분석에도 적용	Bahado-Singh et al (2022a); (2022b)
18	Decision Tree C5.0 (의사결정나무)	- C5.0 알고리즘을 사용하여 의사결정 트리 또는 규칙 세트를 작성. 최대 정보 이득을 제공하는 필드를 기준으로 하여 표본을 분할하는 방식으로 작동 - 범주형 대상만 예측, 범주형 필드가 있는 데이터를 분석하는 경우 노드는 범주를 그룹화	Tran et al(2022)
19	prediction analysis for microarrays	- 여러 분석 과정을 하나의 칩에 집적한 바이오칩 기술의 발달로 인해 DNA microarray 등장 - 짧은 시간에 엄청난 양의 데이터를 생성할 수 있는 가능성을 지니고 있는 도구	Bahado-Singh et al(2023); J Orozco et al (2018)
20	CIBERSORT	- 유전자 발현 데이터로부터 회귀모델 등 전산학적인 계산방법을 이용, 서포트 벡터 회귀 방법 기반의 세포 조성 유추. 클러스터링 알고리즘으로, 미국 스탠포드 대학의 Newman 등에 의해 개발	Yu et al(2021) ;Karisola et al (2021)

참고:

A. M. Newman et. al. (2015). Robust Enumeration of Cell Subsets from Tissue Expression Profiles. Nature Methods, 12(5), 453.DOI: 10.1038/nmeth.3337

Amazon Web Services. (n.d.). *Radial Kernel SVM.* Retrieved Sep 25, 2023 from https://rstudio-pubs-static.s3.amazonaws.com/170893_9e4e88b0dedc4adfa3bd5a87eb64a9ba.html

analystic steps. (n.d.). *A Classification and Regression Tree(CART) Algorithm.* Retrieved Nov 10, 2023 from https://www.analyticssteps.com/blogs/classification-and-regression-tree-cart-algorithm

Analystics Vidhya. (2023.09.21.). *Adaboost Algorithm: Understand Implement and Master AdaBoost*. Retrieved Sep 25, 2023 from https://www.analyticsvidhya.com/blog/2021/09/adaboost-algorithm-a-complete-guide-for-beginners/

ArcGIS Enterprise. (n.d.). *Portal for ArcGIS. Generalized Linear Regression.* Retrieved Nov 10, 2023 from https://enterprise.arcgis.com/en/portal/latest/use/geoanalytics-generalized-linear-regression.htm

Catboost. (n.d.). *CatBoost is a high-performance open source library for gradient boosting on decision trees* Retrieved Sep 25, 2023 from https://catboost.ai/

IBM. (2022). *C5.0 노드.* Retrieved Sep 25, 2023 from https://www.ibm.com/docs/ko/

IBM. (n.d.). *What is the k-nearest neighbors algorithm?.* Retrieved Sep 25, 2023 from https://www.ibm.com/docs/ko/

LightGBM. (n.d.). *Read the Docs. Welcome to LightGBM's documentation!.* Retrieved Sep 25, 2023 from https://lightgbm.readthedocs.io/en/stable/

rdrr.io. (2022.12.28.). *stepwiseCOX: Stepwise Cox Proportional Hazards Regression.* Retrieved Sep 25, 2023 from https://rdrr.io/cran/StepReg/man/stepwiseCox.html

ScienceDirect. (2007). *Logistic Regression.* Retrieved Sep 25, 2023 from *https://www.sciencedirect.com/topics/computer-science/logistic-regression*

Stack Exchange. (2015). *KNN: 1-nearest neighbor*. Retrieved Sep 23, 2023 from https://stats.stackexchange.com/questions/151756/knn-1-nearest-neighbor

TIBCO. (n.d.). *Stepwise Model Builder-Cox Regression Introductory Overview.* Retrieved Sep 25, 2023 from https://docs.tibco.com/pub/stat/14.0.1/doc/html/UsersGuide/user-guide/stepwise-model-builder-cox-regression-introductory-overview.htm

xgboost. (2022). *read the docs. Introduction to Boosted Trees-xgboost 2.0.0.* Retrieved Sep 25, 2023 from https://xgboost.readthedocs.io/en/stable/

이명희. (2002). "혼합 모형을 이용한 microarray 자료의 특이 발현변이 유전자의 추정방법에 관한 연구." 국내석사학위논문 연세대학교 대학원, 서울

ABSTRACT

Current Applications and Future Directions of Artificial Intelligence in Epigenetics

Jiwon Park
Dept. of Medical Law and Ethics
The Graduate School
Yonsei University

Artificial intelligence is indispensable for integrating, interpreting, and managing complex and extensive datasets, playing a pivotal role in decision-making for researchers and clinical settings. Specifically, the application of various artificial intelligence technologies in epigenetics, such as predictive models for determining DNA methylation patterns, not only facilitates epigenetic research but also emerges as a crucial element in providing personalized precision medicine.

From this perspective, this study aims to investigate the extent to which artificial intelligence is applied in the field of epigenetics. A scoping review revealed that artificial intelligence technology is being employed in studies related to epigenetics for purposes such as predicting disease onset, classifying patient groups, and assessing differentiated treatment outcomes. Additionally, a comparison was made using traditional literature review methods to explore the application of AI technology in epidemiology, genetics, and epigenetics. The

results of the traditional literature review indicated that in the fields of epidemiology and genetics, AI models at the 'action' level—automatically performing physical actions or triggering system processes based on inferred results—are actively utilized. However, it was confirmed that such applications have not yet reached this level in the field of epigenetics.

Based on comprehensive results, suggestions for the future direction of applying artificial intelligence in the field of epigenetics were presented, emphasizing the need for the 'enhanced utilization of integrated big data for data cohesion' and the 'expansion of the scope of artificial intelligence technology in the field of epigenetics.' The proposed future direction is expected to enable personalized lifestyle management and a preventive perspective in digital healthcare, contributing to the activation of the healthcare big data platform in South Korea and highlighting the necessity of a national bio big data platform.

However, it is essential to establish the scope of collecting and utilizing epigenetic information for inclusion in the big data platform, and societal discussions regarding legal and ethical regulations are crucial to prevent discrimination issues arising from information collection. With the advancement of such discussions, this study is expected to enable personalized precision medicine. Ultimately, it has the potential to achieve the goal of public health, enabling health management at both individual and group levels.

Key words : Epigenetics, Precision Medicine, Artificial Intelligence, Machine Learning, Future Directions, Genetic Information, Big Data, Omics Data