

범죄자 프로파일링을 위한 성적살인 연구: 랜덤포레스트(Random Forest)를 활용하여

권혁준* · 서종한**

국 | 문 | 요 | 약

본 연구는 국내에서 발생하는 살인범죄 중 성적살인에서 나타나는 특징을 탐색하고 이를 바탕으로 성적살인의 특징을 잘 나타내는 모형을 조사하기 위해 이뤄진 연구이다. 본 연구에서는 기존의 연구들이 가지고 있는 방법론적인 문제를 해결하기 위하여 머신러닝 기반 분석기법인 랜덤 포레스트(random forest)기법을 사용하여 성적살인의 특징을 가장 잘 보여주는 모형을 조사하였고 추가적으로 로지스틱 회귀분석을 활용해 주요 변수들의 방향성을 제시했다. 본 연구에서는 일반적으로 사용되는 가해자 관련 요인 뿐 아니라 범죄행동 관련 요인, 피해자 관련 요인을 모두 투입하여 예측 모형을 탐색했다. 연구결과, 본 연구에서 제시한 성적살인 모형은 어느 정도 용인될 만한 수준의 성과지표를 보고하였으며 데이터가 지속적으로 축적될 필요성이 있는 머신러닝의 특성상 더 높은 성과를 기대할 수 있다. 결과적으로 도출된 중요도가 가장 높은 변수는 가해자와 피해자 간의 관계(비면식)로 나타났다. 로지스틱 회귀분석 결과 가해자의 성별이 남성일수록, 학력이 낮을수록, 미혼일수록 성적살인을 예측하였으며 비면식 관계인 경우, 과잉살해가 없을수록 성적살인일 가능성이 높은 것으로 나타났다. 본 연구는 국내에서 지지 부진한 성적살인에 대한 연구를 확장하고 수사과정에서 도움을 줄 수 있는 범죄 예측 모형을 탐색하였다는 점에 의의가 있고 또한 최근 범죄학에서 소개되고 있는 머신러닝 기법을 성적살인 분석에 적용하였다는 점에서 학문적 시사점을 가진다.

DOI : <https://doi.org/10.36889/KCR.2022.3.31.1.165>.

❖ 주제어 : 성적살인, 랜덤포레스트, 머신러닝, 살인범죄, 범죄 예측

* 영남대학교 심리학과 연구원. 주저자(swj02180@gmail.com)

** 영남대학교 심리학과 교수. 교신저자(jonghansea@yu.ac.kr)

I. 서론

1. 연구의 필요성과 목적

성적살인은 전체적인 발생 비율이 높은 범죄는 아니지만, 그 잔혹성과 충격으로 인해 대중의 큰 사회적 관심을 불러일으키는 범죄이다(Maniglio, 2010). 따라서 성적살인에 대한 연구의 중요성은 높다고 할 수 있다. 하지만 성적살인에 대한 정의 자체가 통일되지 않았을 뿐 아니라 성적살인이 일반적인 다른 살인과 별개로 기록이 되는 경우가 매우 적기 때문에 연구자들이 접근하기 어려운 주제로 꼽힌다(Beauregard et al., 2007). 이러한 요소들은 연구의 결핍을 불러오게 되며, 비단 연구자들 뿐 아니라 현장에서 수사관들의 성적살인에 대한 이해를 저해하기 때문에 수사과정에서 어려움을 야기할 가능성이 높다. 이러한 근거 중 하나로 Berauregard 등(2007), Meloy(2000)와 같은 많은 학자들은 성적살인이 실제로 보고된 수준 혹은 기록된 수준보다 더 많을 수 있다는 점을 지적하고 있다. 이러한 과소평가의 원인 중 하나가 범죄에 대한 이해의 부족일 수 있다. 또한 기존의 성적살인 연구들은 범죄자의 인구통계학적 정보만을 바탕으로 하여 분석한 연구들이 많다(Langevin et al., 1988). 하지만 범죄 수사와 범죄 분석에 있어서 인구통계학적 정보 물론 중요한 요인이라고 할지라도 범죄행동 정보, 범행수법(Modus Operandi)과 관련된 요인도 중요하게 고려할 필요가 있다(Greenall & Richardson, 2015)¹⁾. 특정 범죄에 대한 범죄행동 정보가 축적된다면 범죄자 프로파일링에 있어서 범죄행동 정보를 바탕으로 특정 용의자 군을 선정하여 탐색할 수도 있을 것이다. 즉, 범죄행동 정보라는 사후정보를 통하여 가해자를 예측하는 데 도움을 줄 수 있다.

성적살인에 대한 연구의 부족은 특히 국내에서 두드러지게 나타난다. 성적살인에 대한 연구는 대부분 서구권에 국한되어 연구가 많이 진행되어 왔다(Kubrin, 2003). 이에 반해 국내에서 성적살인에 대한 연구는 거의 없는 수준으로, 권혁준과 서종한(2020)의 연구를 제외하면 국내에서 성적살인을 따로 구분하여 연구의 주제로 선정한 연구가 매우

1) 범행수법(Modes Operandi)이란 범죄 행동시 일어나는 범인의 행동으로, 범인이 남기는 특유의 표식인 시그니처(Signature)와 구분될 필요가 있다. MO의 예로 흉기를 사용한다거나, 시체를 은닉한다거나 하는 등의 행위가 있다.

드문 실정이다. 소수의 성적살인을 언급한 연구(홍성열, 2008; 강은영 등, 2008)에서도 다른 살인범죄들을 종합하여 연구의 대상으로 삼았기 때문에 성적살인 자체에 대한 연구 자체는 적게 수행되었다. 성적살인만을 단일 대상으로 삼아 예측 모형을 탐색한 권혁준과 서종한(2020)의 연구 또한 방법론적 한계로 인해 현실 세계에서 적용할 수 있는 성적살인의 예측 모형을 탐색했다고 하기에는 부족하다. 가장 대표적으로, 해당 연구에서 사용한 의사결정나무 기법은 모델의 용량이 증가하면 증가할수록 컴퓨터가 학습 데이터를 과하게 학습하여 모형의 설명력이 줄어드는 과적합 문제에서 자유롭지 않다(박준영 등, 2016). 그렇기 때문에 데이터가 늘어나게 되면 예측의 오차 또한 커지게 되는 문제가 있다. 하지만 현실세계에서 실제로 발생하는 성적살인은 연구들에서 사용하는 로지스틱 회귀분석이나 한 번의 의사결정나무 기법과 같은 전통적인 통계기법들을 사용하기에는 발생 건수가 많고 관련된 변인들 또한 많다. 따라서 이러한 문제를 해결할 수 있는 머신러닝 기법을 사용하여 실제 데이터에도 활용할 수 있는 예측기(예측 모형)를 탐색할 필요가 있다.

따라서 본 연구에서는 가해자의 인구통계학적 정보뿐 아니라 피해자 관련 요인, 그리고 범죄행동 관련 요인을 모두 고려하여 모형에 투입하고, 이를 통해 머신러닝 기법을 사용하여 빅데이터에 어울리는 예측 모형을 탐색하고자 했다. 또한 기존의 머신러닝 기법의 한계점을 보완하기 위해 새로운 분석기법을 활용하여 보다 과학적이고 범용성 높은 모형을 제시하는데 목적을 두었다.

2. 성적살인

성적살인은 일반적으로 성적 행위가 포함되어 있는 살인범죄를 의미한다. 하지만 성적살인의 정확한 정의는 연구자마다 다르다(Schlesinger, 2007). 가장 대표적으로 Ressler와 같은 연구자들은 성적살인에 대한 준거를 6가지로 제시하였다(Ressler et al., 1988). 첫째, 피해자가 완전 혹은 부분적인 나체상태, 둘째, 피해자 성기 노출, 셋째, 명백히 성적인 시체, 넷째, 시체 특정부위에 물체 삽입, 다섯째, 성적 접촉의 증거 존재, 여섯째, 대체 성행동이나 가학적 성적환상의 표현이 그것이다. 하지만 실제 살인현장에서는 이러한 정보를 쉽게 판별하기 힘들거나 이러한 준거가 존재하지 않는 경우도 많다.

따라서 많은 학자들이 제시한 다양한 견해를 종합해 보았을 때 성적살인이란 가해자의 성적 행동이 나타남과 동시에 사람을 살해하는 행위(Meloy, 2000)라고 볼 수 있다. 따라서 본 연구에서도 Meloy의 정의에 따라 성적살인을 ‘살인 과정에서 성적 행동이 나타난 범죄’를 총칭하여 연구를 진행했다. 그렇기 때문에 본 연구에 포함된 살인범죄는 살인이 1차 동기일 수도 있고 성적행동이 1차 동기일 수 있다. 하지만 자료의 한계로 인해 따로 분류하지 않고 연구를 수행했다. 이러한 성적살인은 남성 가해자가 여성 피해자를 대상으로 행하는 경우가 대부분이다(Robert & Grossman, 1993). 또한 일반적으로 성적살인은 가해자와 피해자가 서로 알지 못하는 사이(비면식 관계)인 경우에 더 많이 발생한다. Healey와 Beauregard(2017)의 연구에서도 많은 성적 살인범들이 자신과 이전에 접촉이 없는 피해자를 선택한다는 결과가 나타난 바 있다.

성적살인은 일반적인 살인범죄, 그리고 일반적인 성범죄와 다른 특성을 가지고 있다. 살인범죄와 성범죄의 동기를 모두 가지고 있고 성범죄 동안 무기를 소유하고 있을 경우에 성적살인으로 발전한다는 견해도 있다(Chan & Heide, 2009). 또한 성범죄 중에서 치명률이 높은 유형이라는 견해 역시 존재한다(Healey & Beauregard, 2017). 종합하자면 성적살인은 일반적인 살인범죄와 일반적인 성범죄와 어느 정도 특성을 공유하면서도 어느 한 범주에 해당하는 범죄라고는 볼 수 없다는 것이다. 따라서 성적살인은 일반적인 살인, 그리고 일반적인 성범죄에 대한 분석과 별개로 따로 연구가 필요한 중요한 범죄 유형이라고 볼 수 있다.

성적살인에 영향을 주는 요소를 살펴보자면, 많은 학자들은 성적살인의 동기 중 성적 판타지를 주된 요소로 꼽는다(McClellan, 2008). 또한 성적 살인범들에게 주로 나타나는 점은 대인관계 속에서 나타나는 문제라고 볼 수 있다. 성적 살인범들은 낮은 자존감을 가지고 있는 경우가 많으며 아동·청소년기의 문제적 가정환경을 가지고 있는 경우가 많다(Brittain, 1970). 또한 성적살인의 동기로 외로움(loneliness)을 꼽는 학자들도 많으며(Grubin, 1994), 성적살인범 중에서 미혼인 범죄자가 많다(Olver et al., 2007)는 연구 결과는 이러한 점을 지지한다. Roberts와 Grossman(1993)의 연구에서도 성적살인범의 약 80%가 미혼이거나 별거, 혹은 이혼 경험이 있어 안정적인 이성 관계를 가지고 있지 않음을 확인했다. 또한 성적 살인범들은 성격장애를 진단받는 경우가 많다²⁾는 결과

2) 반사회성 인격 장애(ASPD)를 포함하여 다양한 성격장애 진단을 나타내었다

를 고려한다면 성격적인 요소가 원인이 될 수 있다.

성적살인을 연구하는 많은 연구자들은 이러한 가해자 관련 요인들뿐만 아니라 범죄현장특성에 대한 고려를 중요하게 생각한다(Ressler et al., 1988). 성적살인범들은 뚜렷한 범행수법을 보인다는 연구결과가 존재한다(Beauregard & Martineau, 2016). 해당 연구에서 성적살인범들은 살인을 한 후 시체를 잘 은닉하지 않거나, 절반 이상이 흥기를 사용하지 않았다. 이러한 결과는 많은 성적살인범들이 살인의 도구로써 칼이나 총과 같은 흥기가 아닌 자신의 신체(주먹 등)를 활용한 개인적인 무기를 주로 사용한다는 다른 연구들에서도 증명된 바 있다(Chan & Heide, 2009; Carter & Holin, 2010). 특히 이러한 개인적인 무기는 성적 만족감을 향상시키기 위한 방법으로 주로 사용된다는 주장도 있다(Schlesinger, 2007). 이러한 살인의 특징은 비조직적(disorganized) 살인의 유형에서 많이 발견되는데 이러한 살인 유형은 과잉살해행동이 많이 나타난다(Ressler et al., 1988).

하지만 이러한 특징적 정보와 별개로 성적살인이 일어나기 위해서는 맥락적 요소가 중요하다는 지적도 있다. Chene과 Cusson(2005)은 비살인 성범죄자와 성적 살인범의 차이에는 맥락적 요소가 중요함을 지적했으며, Chan과 Heide(2009) 또한 성범죄 도중 범죄 현장 주변에 무기가 존재한다면 성적살인으로 발전할 가능성이 높다는 점을 들어 맥락적 요소를 성적살인의 요소로 꼽기도 하였다.

위의 연구들이 성적 살인범들의 특징을 밝히고자 한 연구들이라면, 이러한 성적살인을 분류하고 예측하기 위한 연구들 또한 서구권을 중심으로 이루어진 바 있다. Geberth와 Turco(1997)에 따르면 성적 살인과 비성적 살인을 구분하는 중요한 변수로 피해자와 가해자 간의 면식관계를 주장하기도 하였으며, 사이코패시를 바탕으로 구분하려는 시도도 있었다(Firestone et al., 1998). 국내에서는 권혁준과 서종한(2020)의 연구에서 의사결정나무와 집합 분석을 사용하여 성적살인을 예측하는 모형을 탐색하기도 하였다.

3. 머신러닝 기반 범죄분석

머신러닝은 컴퓨터에게 다수의 데이터를 입력하여 학습을 가능하게 하기 위해 알고리즘과 기술을 개발하는 인공지능의 분야를 의미한다(허선영 등, 2018). 머신러닝에 기반

한 학습방법은 크게 지도학습과 비지도 학습으로 구분된다. 지도학습은 데이터를 입력하고 이에 대한 답을 동시에 입력한 후 컴퓨터가 데이터에 대한 답을 학습하도록 하는 기법이며, 대표적인 예시로 랜덤포레스트, 신경망 분석 등이 존재한다. 이에 반해 비지도 학습은 문제에 대한 답을 실제로 입력하지 않고 컴퓨터가 스스로 데이터간의 차이를 발견하여 이를 학습하는 과정으로 클러스터 분석, 독립 성분 분석 등이 그 예이다.

그중에서 지도학습은 사회과학 분야에서 주로 사용되는 머신러닝 기법으로서 일반적인 사회과학 분야의 데이터에 대한 분석(이기정 & 김영식, 2021; 김영식 등, 2019)뿐만 아니라 범죄학, 범죄심리학 분야에서도 최근에 활용이 증가하고 있는 기법이다(박중영 등, 2016; 김동영 & 정성원, 2021). 특히 범죄학 분야의 경우 CCTV정보를 바탕으로 범죄의 수법을 분류하거나(Bang & Cho, 2016), 시공간 데이터를 활용하여 범죄예측 모형을 비교하고(김동영 & 정성원, 2021), 다양한 빅데이터를 기반으로 범죄의 유형을 예측하는 알고리즘을 개발하는 등(Nasridinov et al., 2013) 국내외에서 공학, 경제학 분야에서 사용하던 머신러닝을 범죄학 연구의 분야에 소개하고, 다양한 연구들을 시도하고 있는 추세이다.

범죄학, 범죄심리학적 측면에서 머신러닝의 활용은 현실세계에서 발생하는 문제점을 해결하는 데 있어서 중요한 역할을 한다. 가장 중요한 점은 범죄는 시민의 안전에 직접적으로 연관되기 때문에 이러한 점이 해소되는 것은 일반 시민들의 불안을 감소시키는데 효과적이라는 것이다(허선영 등, 2018). 특히 살인이라는 범죄는 사람의 생명을 위협하는 범죄이기 때문에 그 위험이 더욱 큰 범죄이고, 따라서 대중들에게 공포감을 불러일으키기 쉬운 범죄 유형이다. 이러한 살인범죄, 특히 연구가 많이 수행되지 않은 성적살인에 대한 예측은 범죄에 대한 위협을 줄이고 시민들의 불안을 감소시키는데 효과적일 것이다. 또한 빅데이터를 활용한 머신러닝 기반 프로그램의 범죄예측 시스템은 이미 국외에서 실시되고 있는 하나의 흐름이라는 점이다. 빅데이터란 정형, 비정형의 대규모 데이터를 관리, 추출, 분석하는 기술 및 자료를 의미하는데, 대표적으로 미국에서는 프레드폴(PredPol), 컴프스탯(COMPSTAT) 등의 머신러닝 기반 범죄예측 프로그램을 사용하여 범죄율을 감소시키는 데 적극적으로 활용하고 있다. 국내에서도 빅데이터를 활용하여 지리적 프로파일링을 통한 범죄수사에 활용하고 있다.

II. 연구방법

1. 연구 대상

본 연구는 경찰청의 과학적 범죄분석 시스템(SCAS)에 2003년부터 2011년까지 입력된 범죄 데이터 중에서 살인범죄와 관련된 데이터를 추출하여 분석하였다. 본 데이터는 경찰청 범죄분석요원이 입력하는 자료로 송치 전 범죄자를 면담한 후 이에 대한 정보를 중 시스템 내 입력하거나 관련된 근거 자료를 저장하게끔 되어 있다. 대표적으로 사회인구학적 특성, 죄중에 따른 범행수법, 범행현장특성, 위험성평가 등 심리검사결과, 범죄분석결과보고서 등이다. 경찰청 범죄분석요원의 협조를 받아 사례를 추출하였고 이에 따라 살인의 유형을 구분하였다. 본 연구에 사용된 비성적 살인(성적살인을 제외한 살인범죄)은 총 360건이었으며 성적살인은 총 94건이었다. 두 종류의 살인범 모두 남성이 여성보다 많았으며, 가해자의 나이는 비성적 살인 가해자의 경우 39.46세였으며 성적살인 가해자는 32.65세였다. 비성적 살인범들은 최종학력이 고등학교 졸업이 많았으며 성적살인의 경우 초등학교 중퇴 이하의 최종학력을 가진 가해자들이 많았다. 범죄기록의 경우 비성적 살인 비해 성적살인범들이 가지고 있는 경우가 많았으나, 구속기록은 두 종류의 살인범 모두 가지고 있지 않은 경우가 많았다. 비성적 살인범들은 기혼상태인 경우가 미혼상태인 경우보다 많은 반면, 성적살인범들은 미혼인 상태가 더 많았다. 구체적인 연구대상자(가해자)의 인구통계학적 정보는 <표 1>과 같다.

〈표 1〉 가해자의 인구통계학적 정보

	비성적살인*	성적살인
가해자 관련 요인		
성별		
남성	320 (88.9%)	93 (98.9%)
여성	40 (11.1%)	1 (1.1%)
연령 M(SD)	39.46 (11.69)	32.65(9.42)
최종 학력		
미교육(초중퇴이하)	37 (10.2%)	33 (35.1%)

	비성적살인*	성적살인
초등학교	68 (18.8%)	20 (21.3%)
중학교	77 (21.4%)	3 (3.2%)
고등학교	140 (38.9%)	28 (29.8%)
대학교 이상	38 (10.6%)	10 (10.6%)
범죄 기록		
있음	246 (68.3%)	75 (79.8%)
없음	114 (31.7%)	19 (20.2%)
구속 기록		
있음	118 (32.7%)	45 (47.8%)
없음	242 (67.3%)	49 (52.2%)
결혼 여부		
미혼	156 (43.3%)	60 (63.8%)
기혼	204 (56.7%)	34 (36.2%)

2. 연구변인

본 연구는 성적살인을 예측하기 위해 다양한 변수들을 투입하여 그 변수들의 영향력을 탐색하고자 하는데 목적이 있다. 따라서 가용할 수 있는 변수들을 최대한 투입하고자 했다. 단, 결측값이 50% 이상 존재하는 변수의 경우 모형에 투입되면 예측력을 저해할 가능성이 있기 때문에(이기정 & 김영식, 2021) 분석에서 제외했다. 또한 최종 학력과 시체 발견 장소와 같은 변인은 원자료에서 매우 많은 범주(10개 이상)를 가지고 있었으나, 범주형 데이터를 랜덤포레스트 기법으로 사용하면 범주가 많을수록 변수 중요도가 높아질 가능성이 있다. 따라서 해당 변인의 경우 최종 학력의 경우 중퇴인 경우 그 이전 수준으로 코딩하였으며, 시체 발견 장소의 경우 건물 내부와 외부로 재코딩하였다.

본 연구에서 사용되는 랜덤 포레스트 기법을 실시하는 분석 라이브러리인 사이킷런은 범주형 변인을 투입할 수 없다는 단점이 존재한다. 따라서 본 연구에서 사용된 범주형 변수들은 모두 원핫코딩(더미변수화)할 수 있는 파이썬 라이브러리(pandas)를 활용하여 전 처리하였으며 총 18개 변수를 투입하여 성적살인에 대한 예측 모형을 탐색했다. 구체적인 변수의 개요는 <표 2>와 같다.

〈표 2〉 랜덤 포레스트 투입 변수의 개요

변인		개요
가해자 관련 요인		
성별	가해자의 성별에 대한 정보	1: 남성, 2: 여성
최종 학력	가해자의 최종 학력 중퇴의 경우 그 이전 학력으로 코딩.	1: 미교육(초중퇴이하) 2: 초등학교(졸이상) 3: 중학교(졸이상) 4: 고등학교(졸이상) 5: 대학교(졸이상)
구속 기록 유무	가해자의 구속 기록 유무	1: 있음, 2: 없음
범죄 기록 유무	가해자의 범죄 기록 유무	1: 있음, 2: 없음
결혼 유무	가해자의 결혼 상태	1: 기혼, 2: 미혼
범죄행동 관련 요인		
살인 장소	살인이 발생한 장소	1: 피해자 관련 장소 2: 가해자 관련 장소 3: 모두 관련된 장소 4: 관련 없음 5: 미확인
과잉살해(overkill)	직접적으로 사망의 원인이 되는 상흔 이외의 치명적 상처가 과도하게 발견된 경우	1: 있음, 2: 없음
범행 시 무기 사용	살인 시 무기를 사용했는지 여부	1: 있음, 2: 없음
예기 사용	범죄 도구로 날카로운 예기(칼 등)를 사용했는지 여부	1: 있음, 2: 없음
둔기사용	범죄 도구로 둔기를 사용했는지 여부	1: 있음, 2: 없음
시체 유기	시체를 범죄현장에서 운반하여 별도의 장소에 유기했는지 여부	1: 있음, 2: 없음
시체 은닉	시체를 범죄 현장에서 은닉하였는지 여부	1: 있음, 2: 없음
시체 발견 장소	시체 발견 장소의 정보	1: 있음, 2: 없음
시체 탈의	시체 발견 당시 시체의 옷이 벗겨져 있었는지 여부	1: 있음, 2: 없음
계획 여부	가해자의 범행 수법 사전 계획성 여부	1: 있음, 2: 없음
피해자 관련 요인		
가해자와의 관계	피해자와 가해자 간의 관계	1: 면식** 2: 일면식*** 3: 비면식
피해자의 성별	피해자의 성별 정보	1: 남성, 2: 여성
동거 여부	피해자와 가해자가 서로 동거를 하는지 여부	1: 예, 2: 아니오

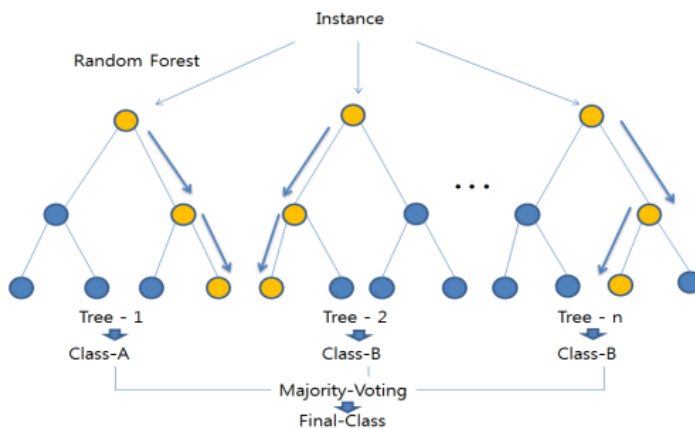
* 통계분석시 모든 범주형 자료는 사이킷런 라이브러리를 통해 원핫코딩; **,***: ‘면식’이란 가족, 친구, 직장동료 등 일상생활에서 지속적인 관계를 맺고 있는 관계를 의미하며, ‘일면식’은 면식관계는 있으나 친밀한 관계가 없는 이웃 등을 의미

3. 통계 분석

본 연구에서는 성적살인의 특징을 가장 잘 나타내는 모형을 탐색하기 위해 머신러닝에서 주로 사용되는 랜덤 포레스트 기법을 이용하여 분석하였다. 랜덤 포레스트 기법은 다수의 의사결정나무(decision tree)를 만들어 반복함으로써 데이터를 분류하고 예측 모형을 선정하는 효과적인 머신러닝 알고리즘 기법이다(유진은, 2015). 즉 의사결정나무를 이용한 앙상블 기법으로, 기존의 의사결정나무에 무작위성을 추가하여 수많은 의사결정나무를 추가하고, 그 중에서 가장 좋은 예측력을 가진 나무의 구성을 탐색하게 된다. 의사결정나무란, 불확실한 상황에서 어떠한 선택(변수의 투입)이 어떠한 결과값을 나타내는지에 대한 결과를 나무 모양으로 나타내는 과정을 의미한다. 이 과정에서 랜덤 포레스트 기법은 투입되는 변수를 무작위로 선정하여 의사결정나무를 만들고 최적의 의사결정나무의 개수를 선정하여 최종 예측기를 생성하는 단계를 거친다. 즉, 가능한 의사결정나무를 모두 비교하여 최적의 나무가 무엇인지를 종합하고 판단하여 최종 모형을 제시하는 기법이다. 변수를 예측하는 모형을 탐색하기 위한 전통적인 통계 분석 기법에는 로지스틱 회귀분석이 있다. 하지만 해당 모형은 변수가 다수 추가되면 자유도가 감소하게 되는 문제가 발생하여 다양한 변수의 투입이 필수적인 빅데이터 기반의 분석에는 적절하지 않다는 단점이 있다. 또한 기존의 의사결정나무 기법은 예측 변수를 평가하고 연관성의 강도와 유의도를 평가하는 기법(권혁준 & 서종한, 2020)이지만 역시 머신러닝 기법에서는 적용하기 어렵다. 하지만 랜덤 포레스트 기법은 의사결정나무를 머신러닝에 적용할 때 나타나는 과적합(학습 데이터에 대해 과하게 학습하여 실제 데이터에 대한 오차가 증가하는 현상; overfitting)문제를 해결할 수 있다는 장점이 있어 주로 사용되는 기법이다(박준영 등, 2016). 따라서 본 연구에는 다양한 요인들을 한 번에 투입하여 랜덤 포레스트 기법을 이용해 성적살인에 대한 모형을 탐색하였으며 그 결과를 예측 모형 성과 지표와 변수 중요도를 통해 제시하였다. 랜덤 포레스트를 포함한 머신러닝 기법에서는 데이터를 학습용과 검증용으로 나누어 분석을 실시하며 일반적으로 7:3 혹은 8:2의 비율로 데이터를 분류한다. 따라서 본 연구에서도 살인범죄 정보 데이터를 학습용 데이터 70%, 검증용 데이터 30%로 나누어 분석을 실시했다. 이러한 작업은 Python 라이브러리의 `train_test_split` 함수를 통하여 무작위로 실시됐다.

랜덤 포레스트를 이용해 도출한 모형의 예측 성과는 정분류율(Accuracy), 정밀도(Precision), 재현율(Recall), F1 점수 등을 통해 평가한다(Dangeti, 2017). 정분류율이란 전체 데이터를 올바르게 분류한 경우의 비율을 의미한다. 따라서 성적살인인 경우 성적살인으로 분류한 확률, 성적살인이 아닌 경우 성적살인이 아니라고 분류한 확률의 합을 의미한다. 정밀도란 신호를 탐지한 경우 중 실제로 신호가 있었던 경우의 비율을 의미한다. 본 사례에서는 성적살인이라고 분류한 경우 중 실제로 성적살인이었던 경우의 비율을 의미한다. 재현율이란 실제로 신호가 있던 경우 중 신호를 탐지한 경우를 의미한다. 본 사례에서는 실제로 성적살인이었던 경우 중 성적살인이라고 올바르게 분류할 확률을 의미한다. F1 점수는 민감도와 재현율의 조화평균을 의미하며 예측 모형의 예측력을 간접적으로 설명해 준다. 이러한 성과지표들은 모두 높을수록 해당 모형이 종속변수를 잘 예측하였다는 것을 나타낸다. 하지만 이러한 성과지표들의 객관적인 평가기준이 따로 존재하지 않기 때문에 기존 연구들과의 비교를 통해 예측력을 비교하는 경향이 있다(홍기혜, 2020). 이해를 돕기 위해 [그림 1]에서 랜덤포레스트의 개념도를 제시하였다.

[그림 1] 랜덤포레스트(이진욱 등, 2017)



추가적으로 모형에 대한 ROC(Receiver Operating Characteristic) 분석을 실시했다. ROC분석은 모형의 성과를 시각적으로 표현하기 위해 사용되었으며 추가적인 모형의 성

과 지표를 나타낼 수 있다. ROC 분석의 경우 ROC curve를 민감도와 1-특이도 쌍³⁾의 도표로 제시하여 해당 모형이 종속변인을 얼마나 잘 예측하는지 평가하는 방법이다(한건환 등, 2015). 여기서 민감도란 신호탐지이론(Greenall, 2012)에 기반하여 신호가 실제로 있는 경우 이를 탐지할 확률을 의미하며, 특이도란 신호가 실제로 없는 경우 이를 탐지하지 않을 확률을 의미한다. 또한 ROC 곡선의 아래 면적을 AUC(Area Under the Curve)라고 하여 수치화하여 제시하게 된다. Simundic(2009)는 AUC를 0.5를 우연 수준으로 하여 총 5가지 수준을 제시하였으며 0.5부터 0.6까지를 나쁨(bad/poor), 0.6부터 0.7까지를 충분함(sufficient), 0.7부터 0.8까지를 좋음(good), 0.8부터 0.9까지를 매우 좋음(very good), 0.9이상을 훌륭함(excellent)으로 판단한다.

본 연구에서는 로지스틱 회귀분석을 추가로 실시하여 랜덤 포레스트에서 도출된 변수들 중 상위변수를 추출하여 변수들의 방향성을 살펴보고자 하였다. 랜덤 포레스트 기법은 변수들의 관계를 해석하기 어렵다는 단점이 있기 때문에 추가적인 보완방식이 필요한데 로지스틱 회귀분석이 그 중 하나의 방식으로 사용된다(이기정 & 김영식, 2021). 따라서 랜덤 포레스트 기법이 변수들의 중요성을 알아보기 위한 방법이라면, 로지스틱 회귀분석의 경우 중요하다고 판단된 변수들의 방향성을 알아보기 위한 추가적 분석 방법이라고 이해하면 될 것이다.

모든 분석은 Python 3의 사이킷런(scikit-learn; 이하 sklearn) 라이브러리를 통해 실시했다. 해당 라이브러리는 머신러닝에 사용되는 대표적인 도구이다.

Ⅲ. 연구결과

랜덤 포레스트 기법을 활용하여 성적살인에 대한 탐색 모형을 선정하였다. <표 3>은 모형의 성과를 나타낸다. 모형의 성과는 정분류율, 정밀도, 재현율, F1 점수를 통해 검증하였다. 해당 예측기의 경우 학습과정에서 1.0으로 매우 좋은 분류율을 나타냈으며 학습 이후 실제 데이터를 활용하여 검증하는 과정에서도 0.81로 높은 수준의 분류율을 보여

3) 연구자에 따라 정반응률(true positive rate)과 거짓 양성률(false positive rate)의 쌍으로 나타내기도 한다. 각각 민감도와 1-특이도의 의미와 동일하다.

해당 모형이 살인의 종류를 올바르게 예측할 확률이 높았음을 확인할 수 있다. 본 예측기의 경우 정밀도는 0.71로 나타났다. 해당 예측기의 재현율은 .64이었으며, 마지막으로 해당 예측기의 F1점수는 .67로 어느 정도 보통 수준 이상으로 측정되었다.

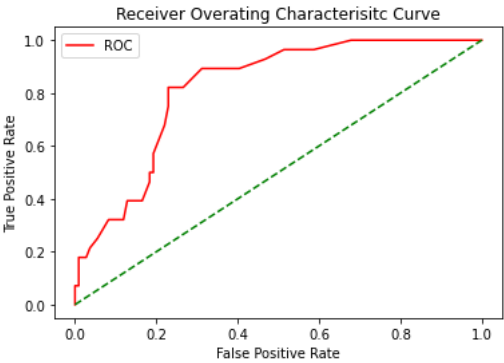
〈표 3〉 탐색 모형의 성과지표

성과지표	수치
정분류율(Accuracy)*	0.81
정밀도(Precision)	0.71
재현율(Recall)	0.64
F1-점수(F1-score)	0.67

* 모형의 최종 정분류율을 의미

다음으로 모형의 예측력을 시각적으로 확인하기 위해 ROC 분석을 실시하였다. ROC 분석의 경우 곡선의 아래쪽 면적이 넓을수록 예측력이 높다고 판단하며, 이를 AUC라는 수치로 표현한다. 본 연구에서 탐색한 예측 모형의 AUC는 0.71로 ‘ 좋음 ’ 정도의 예측력을 나타내었다. 구체적인 ROC곡선의 모습은 [그림 2]와 같다.

[그림 2] 탐색 모형의 ROC curve



* X-axis: 1-특이도, Y-axis: 민감도

랜덤 포레스트 기법은 모형 자체의 설명력을 보여줄 뿐 아니라 투입된 변인들의 설명력, 즉 변수의 중요도(feature importance)를 보여준다. 변수 중요도는 모형의 설명력을 총합 1로 가정했을 때 각 변수들이 가지는 중요도를 의미한다. 따라서 변수 중요도는 0부터 1까지의 값을 가지며 1에 가까울수록 높은 중요도를 가지고 있음을 나타낸다. 본 연구는 범주형 데이터를 더미변수로 변환하여 분석하였으므로 각 더미에 대한 변수 중요도가 측정되었다. 더미를 합산하여 중요도를 측정한 결과 투입 변인 중 가장 높은 중요도를 가진 변인들을 상위 5개 추출하면 가해자·피해자간 관계, 가해자의 최종 학력, 살인 장소, 가해자의 성별, 피해자의 성별로 나타났다. 하지만 사이킷런 모듈을 이용한 랜덤포레스트는 범주형 자료를 원핫코딩(1과 0으로 이루어진 더미변수화)하여 사용하므로 범주가 많은 변인의 중요도가 높아지는 단점이 있다. 본 자료에서는 상위 3개의 변수가 그러한 특성을 가지고 있다. 따라서 추가적으로 3개의 변수를 추출하자면 과잉살해(overkill), 가해자의 구속기록, 결혼 여부가 상위 중요도를 차지하는 것으로 나타났다. 구체적인 변수 중요도의 값은 <표 4>, [그림 3]과 같다.

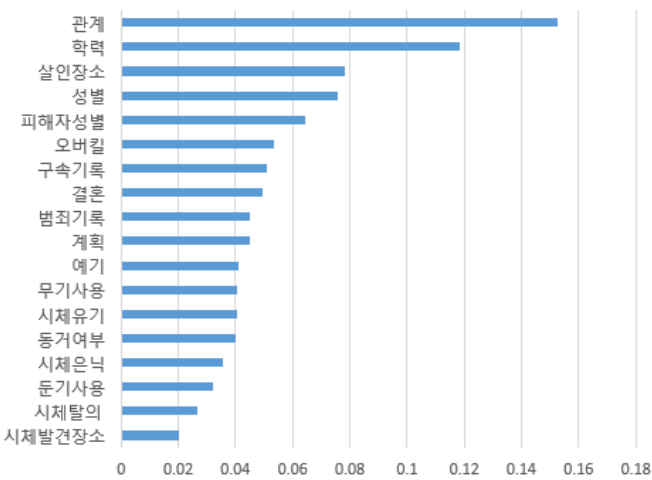
<표 4> 예측 모형의 변수 중요도(feature importance)

변인	중요도
가해자 관련 요인	
최종 학력	0,1183
성별	0,0758
구속 기록 유무	0,0508
결혼 여부	0,0496
범죄 기록 유무	0,0450
범죄행동 관련 요인	
살인 장소	0,0781
과잉살해(overkill)	0,0536
계획 여부	0,0450
예기 사용	0,0411
범행 시 무기 사용	0,0404
시체 유기	0,0403
시체 은닉	0,0356
둔기 사용	0,0322
시체 탈의	0,0264

변인	중요도
시체 발견 장소	0,0203
피해자 관련 요인	
가해자와의 관계	0,1526
피해자의 성별	0,0642
동거 여부	0,0401

*각 요인별로 내림차순으로 정렬

[그림 3] 예측 모형의 변수 중요도 도표



* 전체 요인에 따라 내림차순으로 정렬(관계>학력>살인 장소 순)

다음은 추가적으로 위에서 추출했던 8가지 상위 중요변수(가해자 관련 변인: 성별, 최종학력, 결혼 여부, 범죄기록 유무; 피해자 관련 변인: 성별, 가해자와의 관계; 범죄행동 관련 변인: 살인 장소, 과잉살해)들의 예측방향을 측정하기 위해 로지스틱 회귀분석을 실시했다. 분석 결과 가해자 성별이 남성일수록, 가해자의 최종 학력이 교육을 받지 못했거나 초등학교, 고등학교일수록, 가해자가 범죄 이력이 있을수록 약간의 면식 혹은 비면식 관계일수록, 살인 장소가 피해자 혹은 가해자 중 한 명과 관련이 있을수록 성적살인일 가능성이 높았다. 대부분의 변수는 통계적으로 유의미하게 성적살인을 예측 하였으나 최종 학력의 경우 교육을 받지 않은 경우를 제외하고는 성적살인을 예측하지 못했다. 구

체적인 분석 결과는 <표 5>와 같다.

〈표 5〉 상위 8개 변수에 대한 로지스틱 회귀분석

변수		회귀계수(B)
가해자 성별	남성	2.17***
	미교육(중퇴이하)	1.367**
가해자 최종 학력	초등학교	.482
	중학교	-.216
	고등학교	.483
가해자 결혼 여부	기혼	-.325*
가해자 범죄 기록 유무	있음	.339*
피해자 성별	남성	-.654***
	면식	-1.541**
가해자-피해자 간 관계	일면식	.115
	피해자 관련	18.46***
살인 장소	가해자 관련	18.01***
	피해자와 가해자 모두 관련	17.04***
	확인되지 않음	-17.58***
과잉살해(overkill)	있음	-.261*

* p<.05, **p<.01, ***p<.001.

! 참조: 가해자 성별(여성), 가해자 최종 학력(대학교 이상), 가해자 결혼(미혼), 가해자 범죄기록(없음), 피해자 성별(여성), 가해자-피해자 관계(비면식), 살인 장소(피해자와 가해자 관련 없음), 과잉살해(없음)

IV. 논의

본 연구는 다수의 변인들을 종합적으로 고려하여 현실상황에 맞는 성적살인의 특징을 잘 나타내는 모형을 탐색하고자 하는 목적으로 실시됐다. 따라서 총 18가지의 가해자 관련 요인, 피해자 관련 요인, 범죄행동 관련 요인을 모두 투입하여 머신러닝 기법을 활용하여 성적살인의 특징을 가장 잘 보여주는 요인을 찾고자 했다. 랜덤포레스트와 ROC 기법을 통하여 모형을 탐색한 결과 정분류율은 .81수준으로 높은 수준으로 나타났으며

정밀도와 재현율도 0.71, 0.64 정도로 사회과학 분야에서 수용될 만한 정도의 성과 지표를 나타내었다. Kline(2016)에 따르면 정분류율은 .70이상인 경우 좋은 예측성고를 나타내는 모형이라 하였으며, 전체 모형의 성과를 나타내는 F1 점수 또한 0.67로 좋은 수준은 아니지만 어느 정도 수용할 만한 정도임이 확인됐다. 민감도와 재현율의 경우 정확한 기준점은 존재하지 않지만, 일반적으로 사회과학 분야에서는 60%에서 70% 이상의 성과라면 용인하는 경향이 있다. 랜덤포레스트 기법을 통해 도출되는 특이도의 경우 민감도와 트레이드 오프(trade-off)관계에 있는 변인이기 때문에 높은 민감도의 예측 모형의 경우 낮은 특이도를 보고한다(유진은, 2015). 따라서 본 연구에서는 특이도를 모형의 성과 판단에 특별하게 고려하지 않았다. 또한 ROC분석에서 나타난 AUC값은 .71 정도로 ‘타당함’ 수준으로 나타나고 있다는 점을 고려하면 해당 모형은 어느 정도 받아들여질 만한 수준의 성과를 보았다고 판단된다. 또한 머신러닝을 사용한 알고리즘은 장기간의 데이터 축적을 필요로 하기 때문에(Choi & Park, 2015), 이후에 데이터가 더 축적된다면 모형의 예측력이 더 높아질 가능성이 있다고 볼 수 있다.

구체적인 변수 중요도 분석 결과에 따르면, 상위 5개의 설명변수는 ‘피해자-가해자 간 관계’, ‘가해자의 최종 학력’, ‘살인 장소’, ‘가해자의 성별’, ‘피해자의 성별’이었다. 이러한 결과는 성적살인을 예측하기 위해서는 가해자와 피해자의 관계가 중요하다는 것을 보여준다. 이는 성적살인의 경우 피해자와 가해자 간에 비면식 관계에서 성적살인이 대부분 나타난다는 선행 연구들과 같은 결과이다. 따라서 성적살인을 분석할 때 피해자의 주변인과의 관계를 고려하는 것이 중요하다는 결과를 보여준다. 또한 전통적으로 살인범죄와 높은 연관성이 발견된 가해자 관련 요인들이 본 연구에서도 마찬가지로 높은 중요도를 가지고 있는 것으로 측정됐다. 살인범들의 경우 낮은 학력을 가지고 있는 경우가 많으며, 특히 성적살인의 경우 대부분의 범죄자가 남성이라는 결과(Beauregard & Martineu, 2016)가 본 연구에서도 재현됐다. 또한 살인행위가 발생한 장소 역시 선행연구에서 나타난 결과와 유사한 결과가 나타났다. 권혁준과 서종한(2020)의 연구에서도 피해자와 가해자 모두 관련이 있거나 피해자와 관련이 있는 장소에서 성적살인이 많이 벌어지는 것으로 보고하고 있다. 이러한 점은 일상 활동이론(Cohen & Felson, 1979)(routine activity theory)에 기반하여 이해할 수 있을 것이다. 일상 활동이론에 따르면 범죄가 발생하는 요인으로 동기화된 범죄자, 적절한 피해자, 보호수단의 부재의 세

가지 요소가 결합을 주장하였다. 이러한 관점에서 살인이 벌어진 장소를 해석해본다면, 가해자와 관련된 장소는 보호수단의 부재를 확인할 수 있는 기회가 다른 장소에 비해 높다는 특징이 있다. 이는 ‘보호수단의 부재’와 관련된 요인으로 해석할 수 있다. 또한 피해자와 관련된 장소는 일반적으로 안정감을 느끼거나 방심을 할 가능성이 높다는 특징이 있으므로 ‘적절한 피해자’와 관련된 요인을 보여준다. 따라서 이러한 내용은 성적살인의 특징이라기보다는 일반적으로 일어나는 범죄의 특성을 어느 정도 반영한다는 것으로 해석해도 좋을 것이다.

살펴본 바와 같이 전체적으로 가해자 관련 정보가 높은 중요도를 보이는 것이 확인되었으나 Python을 활용한 랜덤 포레스트 기법의 단점으로 범주가 많을수록 중요도가 낮게 나온다는 점을 고려한다면, 하위 순번에 있는 변인에 대한 중요성을 무시할 수 없다. 따라서 살인 장소, 피해자의 성별 등 가해자 요인 뿐 아니라 범죄행동 관련 요인이나 피해자 관련 요인도 성적살인을 예측하는 데 중요한 변수라는 것을 의미한다. 범죄행동 관련 요인에서 이분형(binary) 변수로 코딩된 요인 중에서 가장 높은 중요도를 보고한 것은 과잉살해(overkill)이다. 과잉살해란 일반적으로 직접적으로 사망의 원인이 되는 상흔 이외의 치명적 상처가 발견된 경우를 의미한다(Sea & Beauregard, 2021). 이러한 과잉살해는 일반적으로 면식관계에서 벌어진 살인범죄에서 주로 발견된다(Martins, 2019). 면식범에 의해 벌어진 살인이 비면식범에 의해 벌어진 살인보다 폭력적이라는 주장도 존재한다(Last & Fritzson, 2005). 권혁준과 서종한(2020)의 연구에서도 성적살인에서 면식과 비면식을 구분하는 가장 주요한 변인 중의 하나로 과잉살해의 존재를 꼽은 바 있다. 이러한 높은 관련성으로 인해 해당 모형에서 과잉살해의 중요도가 높이 측정된 것으로 보인다. 무기의 사용은 비교적 낮은 중요도를 나타내었다. 이러한 결과는 성적 살인범들은 주로 개인적인 가해방식(예, 주먹으로 때리는 등)을 사용하거나 도구적으로 무기를 사용하지 않는다는 연구(Carter & Hollin, 2010)와 유사한 결과라고 볼 수 있다. 또한 시체 유기 및 은닉 또한 마찬가지로 낮은 중요도를 보고하였는데, 이는 기존에 수행된 권혁준·서종한(2020)의 연구와 동일한 결과이다.

본 연구에서는 추가적으로 중요도 상위 변수들의 방향성을 살펴보기 위하여 로지스틱 회귀분석을 실시하였다. 로지스틱 분석의 경우 범주의 개수에 따른 중요도 추정 문제를 고려하여 상위 5개 변수가 아닌 상위 8개 변수에 대하여 분석을 실시했다. 이러한 문제

점은 분석을 하기 위해 사용된 sklearn 라이브러리의 단점이기 때문에 학습 과정에서는 고려하지 않았다. 로지스틱 회귀분석의 결과도 예측 모형과 마찬가지로 선행연구와 동일하게 나타났다. 구체적으로 종합하자면, 가해자 요인을 고려하면 남성 범죄자가 성적살인 범죄를 많이 일으키며, 가해자는 최종 학력이 낮을 가능성이 높고, 범죄 기록이 있을 가능성이 높았다. 피해자 요인을 고려하면 피해자의 성별이 여성인 경우 성적살인일 가능성이 높았으며 가해자와 피해자 간 면식이 없을수록 성적살인일 가능성이 높았다. 범죄행동 관련 요인을 고려하면 살인 장소가 가해자 혹은 피해자와 관련이 있을 가능성이 더 높았고, 과잉살해 행동은 많이 일어나지 않았음을 확인할 수 있다. 이에 반해 가해자의 학력은 교육을 받지 않은 경우를 제외한 나머지의 경우는 성적살인을 예측하는 데 통계적으로 유의미한 수준을 만족하지 못했다. 따라서 이에 의무교육과정을 이수하지 못한 범죄자들의 경우를 제외하고는 교육 수준은 성적살인을 예측하는데 크게 작용하지 않는 것으로 보인다. 즉, 랜덤 포레스트 기법을 통해 도출된 가해자의 최종 학력 변수의 높은 변수 중요도는 모든 범주에 따른 결과라기보다는 교육을 받지 못한 범죄자들의 데이터에 의해 나타났다고 해석할 수 있을 것이다. 범죄에 대해 연구한 선행연구들 또한 낮은 교육 수준의 범위를 정확하게 한정하지 않았으며, 나라마다 교육 수준의 범위도 천차만별이기 때문에(권혁준 등, 2020; Groot & van den Brink, 2010; Hjalmarsson & Lochner, 2012) 초등학교 이상의 최종 학력이 성적살인을 예측하지 못한 것은 이러한 원인 때문일 수 있다.

추가된 3개의 변인 중 가해자의 결혼 여부 역시 기존에 수행된 성적 살인범들의 특징을 동일하게 보고하고 있다. Oliver 등(2007)의 연구에 따르면 성적 살인범들은 대부분 사회적 고립상태에 있으며 외로움을 많이 보고한다. 또한 대부분 결혼한 적이 없거나 혹은 혼자 거주하고 있는 경우가 많다는 점을 고려한다면 본 연구의 결과는 받아들여질 수 있다고 판단된다. 또한 앞서 살펴본 것과 마찬가지로 과잉살해가 발견되지 않은 경우가 성적살인을 더 예측하였음을 확인할 수 있다. 마지막으로 가해자의 구속 이력은 이전의 선행연구에서는 주로 발견되지 않은 특성이지만 일반적으로 범죄 이력 혹은 구속 이력은 범죄 발생에 있어 중요한 변인으로 고려된다는 점(권혁준 & 서종한, 2020b; Hillbrand et al., 1999; Bourget et al., 2007)을 고려한다면 이러한 점이 예측 모형의 중요도에 영향을 미쳤다고 해석할 수 있을 것이다.

본 연구는 몇 가지의 시사점을 제공할 수 있다. 첫 번째로, 국내에서 지지부진한 성적 살인에 대한 연구를 확장하였다는 점이 중요한 의의일 것이다. 성적살인은 전체 살인범죄의 약 5%정도로 비율이 높지는 않은 범죄임에도 불구하고(Greenall, 2012) 범죄의 심각성과 사회적 관심은 매우 높은 범죄이다(Chan et al., 2010). 따라서 국내의 학계와 사회적인 관심과의 괴리상태가 지속되고 있는 실정이다. 하지만 본 연구는 성적살인에 대한 예측변수를 과학적 방법으로 탐색함으로써 성적살인에 대한 지식을 확장하고 후속 연구들을 촉진할 수 있다는 점에서 학문적으로 중요한 의의를 가진다. 둘째, 성적살인에 대한 연구를 통해 국내 범죄수사의 효율성을 향상시키는데 도움을 줄 수 있을 것이다. 경찰력의 증가는 범죄의 억제에 중요한 역할을 하는 것으로 보고되고 있다(Evans & Owens, 2007). 하지만 국내의 치안을 담당하는 경찰력은 경찰 1인당 492명의 인구를 담당해야 할 정도로 부족한 실정이다(양승돈, 2013). 하지만 국가의 경제적, 제도적 고려사항으로 인해 단순히 경찰력을 단기간에 증가시키기에는 무리가 있다. 따라서 현시점에서 중요한 것은 ‘부족한 경찰력을 어떻게 효율적으로 활용할 것인가’와 관련된 내용일 것이다. 실제 범죄수사를 함에 있어서 범죄의 발생과 유지의 원인, 변수들의 관계 등 소위 수사지식(investigative knowledge)은 수사의 효율성을 높이는 데 중요한 요인이다(Greenall & Richardson, 2015). 본 연구를 시작으로 성적살인에 대한 지식이 쌓인다면 수사지식을 형성하는 데 중요한 역할을 할 수 있고 이는 경찰력의 효율성 증대에도 긍정적인 영향을 미치리라 기대한다. 실제로 Schlesinger(2007)는 연구의 축적이 효율적인 경찰수사에 도움을 줄 것이라고 주장하였다. 셋째, 수사지식과 관련한 의의점을 추가적으로 생각해 볼 수 있다. 수사지식은 매우 포괄적인 개념을 의미하고 있다. 즉 기존에 범죄분석에 활용하였던 가해자들의 인구통계학적인 정보 또한 수사지식의 하나의 예이다. 하지만 이러한 요인뿐만이 아니라 범죄현장 혹은 범죄행동에서 발견할 수 있는 요인들과 피해자 관련 요인들 또한 수사지식에서 중요한 위치를 가지고 있다(Greenall & Richardson, 2015). 예를 들어, 범죄현장에서 발견된 정보들을 바탕으로 가해자의 특성을 예상해 볼 수 있다. 본 연구에서의 결과를 참조한다면 과잉살해 행동이 나타난 성적 살인 현장의 정보를 바탕으로 가해자가 ‘피해자와 비면식관계인 교육을 받지 않은 남성’이라는 점에 중점을 두고 용의자를 탐색할 수 있을 것이다. 따라서 본 연구에서는 성적 살인을 예측하기 위한 모형을 탐색하는 과정에서 가해자 관련 변수들 뿐 아니라 범죄행

동 관련 변수와 피해자 관련 변수를 모두 모형에 투입하여 예측력을 탐색하였다. 넷째, 국내외에서 사용하는 범죄분석에 대한 기존의 방법론을 개선하였다는 점이 있다. Langevin(1968)에 따르면 기존의 범죄연구는 대부분 가해자의 인구통계학적 정보를 활용한 기본적인 수준의 연구에 그치고 있다. 최근에 와서야 과학적인 기법들을 활용하여 범죄수사에 활용하고자 하는 움직임이 있으나 그 수가 아직까지 충분하지 않고, 이기정과 김영식(2021)이 제기한 의사결정나무 기법의 한계점 등을 고려하지 않고 현실에서 사용하기 어려운 기법들을 활용하는 연구들도 많이 존재한다. 따라서 본 연구는 그러한 문제점을 해소할 수 있는 머신러닝 기반의 랜덤 포레스트 기법을 활용하여 보다 과학적으로 성적살인에 대한 탐색적 모형을 발견하고자 했다.

하지만 동시에 본 연구는 몇 가지 한계점도 가지고 있다. 가장 첫 번째로는 머신러닝을 기반한 연구를 수행하였으나 빅데이터를 바탕으로 하는 머신러닝 기법에 비해서 작은 데이터셋의 크기를 활용하였다는 것이다. 물론 개인정보 등의 문제로 인해 대규모의 데이터를 활용하지 못하였기 때문에 나타난 한계점이지만, 후속연구에서는 보다 많은 데이터를 포함하여 모형에 투입할 필요가 있다. 자료의 한계점으로 유경진 등(2019)의 연구에서 지적하였듯이, SCAS는 범죄의 정보를 정확히 반영하지 못한다. 해당 연구에서는 이러한 문제에 대한 정책적, 조직적 해결방안을 제시하였으나, 개인 연구자의 경우 그러한 방법은 불가능하다. 따라서 후속 연구에서는 보다 많은 정보를 활용할 수 있는 자료(수사 보고서 등)를 제공할 수 있는 기관과 협력하여 진행할 필요가 있다. 또한 기존에 수행된 연구 결과들에 대한 메타적 접근을 통해 교차 검증하는 것도 하나의 방식이 될 것이다. 두 번째는 범죄를 억제하는 것과 마찬가지로 범죄의 재범을 막는 것도 중요한 측면이지만, 범죄의 동기나 가해자의 성격특성과 같은 데이터는 활용하지 않았기 때문에 본 연구의 예측 모형은 성적살인 자체를 예측하는 데 있어서는 효율적일 수 있으나, 성적살인의 재범을 억제하기 위해서는 그 활용도가 떨어진다는 점이다. 따라서 후속 연구에서는 재범 방지를 위한 또 다른 모형을 탐색할 필요가 있다. 마지막으로 범죄수사 과정에서 빅데이터의 활용 가능성에 대한 의문점이 존재한다. 권양섭(2017)은 범죄예방 및 수사에 있어서 시스템 구축의 한계점을 제시한 바 있다. 가장 중요한 점은 개인정보의 활용과 관련된 부분으로, 범죄수사를 위함이라고 할지라도 원자료에 대한 정당한 접근권이 필요하다는 것이다. 이 외에도 국민의 기본권 침해, 범죄혐의 인지 이전의 데이터

분석등과 관련된 문제에서 법적 문제를 일으킬 수 있다고 주장한다. 따라서 본 연구와 같이 빅데이터를 활용한 범죄예측 연구가 실제로 범죄 억지에 역할을 한다는 근거가 계속해서 제시되어야 할 것이다.

참고문헌

1. 단행본

- Beauregard, E., & Martineau, M. (2016). *"The sexual murderer: Offender behaviour and implications for practice"*. Routledge.
- Dangeti, P., 2017, *Statistics for Machine Learning*. Birmingham, UK: Packt Publishing Ltd.
- Kline, R. B. (2016). *Methodology in the social sciences. Principles and practice of structural equation modeling (4th ed.)*, Guilford Press.
- Martins, T. (2019). *"Overkill: a first run definition"*, Doctoral dissertation, University of Central Florida Orlando, Florida.
- McClellan, J. (2008). "Sexual (Lust) Homicide", In *Serial murder and the psychology of violent crimes* (pp. 229-244). Humana Press.
- Nasridinov, A., Ihm, S. Y., & Park, Y. H. (2013). "A decision tree-based classification model for crime prediction", In *Information Technology Convergence* (pp. 531-538). Springer, Dordrecht.
- Ressler, R. K., Burgess, A. W., & Douglas, J. E. (1988). *Sexual homicide: Patterns and motives*, Simon and Schuster.

2. 논문

- 강은영, 박형민, 박현수, 조소연. (2008). "살인범죄의 실태와 유형별 특성: 연쇄살인, 존속살인 및 여성살인범죄자를 중심으로", 『형사정책연구원 연구총서』, 1-2.
- 권양섭. (2017). "범죄예방과 수사에 있어서 빅데이터 활용과 한계에 관한 연구", 『법학연구』, 17(1): 179-198.
- 권혁준, 김성혜, 서종한. (2020). "한국 방화범죄의 유형과 연쇄 방화의 예측요인", *Journal of The Korean Data Analysis Society*, 22(1): 387-401.
- 권혁준, 서종한. (2020). "범죄자 프로파일링을 위한 성적살인 연구:-피해자-가해자 관계

- 를 중심으로”, 『교정담론』, 14(2): 163-192.
- 권혁준, 서종한. (2020b). “범죄자 프로파일링을 위한 존속 살인 연구”, *Journal of The Korean Data Analysis Society (JKDAS)*, 22: 2153-2169.
- 김동영, 정성원. (2021). “시·공간 데이터를 활용한 머신러닝 기반 범죄예측모형 비교”, 『대한건축학회논문집』, 37(1), 135-143.
- 김영식, 이은정, 주현준. (2019). “랜덤 포레스트를 활용한 대학 수시 전형 선택 관련 예측 요인 탐색”, 『교육과학연구』, 50(4): 233-255.
- 박준영, 채명수, 정성관. (2016). “실시간 범죄 예측을 위한 랜덤포레스트 알고리즘 기반의 범죄 유형 분류모델 및 모니터링 인터페이스 디자인 요소 제안”, 『정보과학회 컴퓨팅의 실제 논문지』, 22(9): 455-460.
- 양승돈. (2013). “언어 네트워크 분석 (Semantic Network Analysis) 을 활용한 민간경비 분야의 연구 경”, 『한국콘텐츠학회논문지』, 13(11): 894-901.
- 유경진, 김애경, 이정호, 정재봉. (2019). “변사사건 분석을 통한 과학적범죄분석시스템 (SCAS)의 문제점과 개선방안”, 『과학수사학회지』, 13(3): 173-179.
- 유진은. (2015). “랜덤 포레스트”, 『교육평가연구』, 28: 427-448.
- 이기정, 김영식. (2021). “랜덤 포레스트 기법을 활용한 장애인로자의 고용유지 예측 요인 탐색”, 『특수교육논총』, 37(2): 73-95.
- 이진욱, 유국현, 문병민, 배석주. (2017). “감성분석과 Word2vec을 이용한 비정형 품질 데이터 분석”, 『품질경영학회지』, 45(1): 117-127
- 한건환, 민윤기, 이장규. (2015). “성범죄자 분류 기준에 따른 시각반응시간과 선택반응시간에 대한 ROC 분석”, 『사회과학연구』, 26(3): 47-62.
- 허선영, 김주영, 문태현. (2018). “머신러닝기반 범죄발생 위험지역 예측”, 『한국지리정보학회지』, 21(4): 64-80.
- 홍기혜. (2020). “랜덤포레스트 머신러닝 알고리즘 기반 남·여 청소년의 자살생각 예측 및 분석”, 『한국사회복지학』, 72(3): 157-180.
- 홍성열. (2008). “살인 범죄자 프로파일링”, 『사회과학연구』, 47(1): 227-254.
- Bang, S. H., & Cho, H. B. (2016). “Learning Method for Real-time Crime Prediction Model Utilizing CCTV”, *Journal of The Korea Society of*

- Computer and Information*, 21(5): 91-98.
- Beauregard, E., Proulx, J., Rossmo, K., Leclerc, B., & Allaire, J. F. (2007). "Script analysis of the hunting process of serial sex offenders", *Criminal justice and behavior*, 34(8): 1069-1084.
- Bourget, D., Gagné, P., & Labelle, M. E. (2007). "Parricide: a comparative study of matricide versus patricide", *Journal of the American Academy of Psychiatry and the Law Online*, 35(3): 306-312.
- Brittain, R. P. (1970). "The sadistic murderer", *Medicine, Science and the Law*, 10(4): 198-207.
- Carter, A. J., & Hollin, C. R. (2010). "Characteristics of non-serial sexual homicide offenders: A review", *Psychology, Crime & Law*, 16(1-2): 25-45.
- Chene, S., & Cusson, M. (2005). "Sexual Murderers and Sexual Aggressors: Intention and Situation" (From Sexual Murderers: A Comparative Analysis and New Perspectives, P 71-86, 2005, Jean Proulx, Eric Beauregard, et al., eds.--See NCJ-220255).
- Chan, H. C., & Heide, K. M. (2009). "Sexual homicide: A synthesis of the literature", *Trauma, Violence, & Abuse*, 10(1): 31-54.
- Choi, D. H., & Park, J. O. (2015). "The Application Method of Machine Learning for Analyzing User Transaction Tendency in Big Data environments", *Journal of the Korea Institute of Information and Communication Engineering*, 19(10): 2232-2240.
- Cohen, L. E., & Felson, M. (1979). "Social change and crime rate trends: A routine activity approach", *American sociological review*, 588-608.
- Evans, W. N., & Owens, E. G. (2007). "COPS and Crime", *Journal of public Economics*, 91(1-2): 181-201.
- Geberth, V. J., & Turco, R. N. (1997). "Antisocial personality disorder, sexual sadism, malignant narcissism, and serial murder", *Journal of Forensic*

- Science*, 42(1): 49-60.
- Greenall, P. V. (2012). "Understanding sexual homicide", *Journal of sexual aggression*, 18(3): 338-354.
- Greenall, P. V., & Richardson, C. (2015). "Adult male-on-female stranger sexual homicide: A descriptive (baseline) study from Great Britain", *Homicide Studies*, 19(3): 237-256.
- Groot, W., & van den Brink, H. M. (2010). "The effects of education on crime", *Applied economics*, 42(3): 279-289.
- Grubin, D. (1994). "Sexual murder", *The British Journal of Psychiatry*, 165(5): 624-629.
- Healey, J., & Beauregard, E. (2017). "Impulsivity as an etiological factor in sexual homicide", *Journal of Criminal Justice*, 48: 30-36.
- Hillbrand, M., Alexandre, J. W., Young, J. L., & Spitz, R. T. (1999). "Parricides: characteristics of offenders and victims, legal factors, and treatment issues", *Aggression and Violent Behavior*, 4(2): 179-190.
- Hjalmarsson, R., & Lochner, L. (2012). "The impact of education on crime: international evidence", *CESifo DICE Report*, 10(2): 49-55.
- Iqbal, R., Murad, M. A. A., Mustapha, A., Panahy, P. H. S., & Khanahmadliravi, N. (2013). "An experimental study of classification algorithms for crime prediction", *Indian Journal of Science and Technology*, 6(3): 4219-4225.
- Kubrin, C. E. (2003). "Structural covariates of homicide rates: Does type of homicide matter?", *Journal of Research in Crime and Delinquency*, 40(2): 139-170.
- Langevin, R., Ben-Aron, M. H., Wright, P., Marchese, V., & Handy, L. (1988). "The sex killer", *Annals of Sex Research*, 1(2): 263-301.
- Last, S. K., & Fritzon, K. (2005). "Investigating the nature of expressiveness in stranger, acquaintance and intrafamilial homicides", *Journal of Investigative Psychology and Offender Profiling*, 2(3): 179-193.

- Maniglio, R. (2010). "The role of deviant sexual fantasy in the etiopathogenesis of sexual homicide: A systematic review", *Aggression and Violent Behavior*, 15(4): 294-302.
- Marshall, W. L., & Barbaree, H. E. (1990). "An integrated theory of the etiology of sexual offending", In *Handbook of sexual assault* (pp. 257-275). Springer, Boston, MA.
- Meloy, J. R. (2000). "The nature and dynamics of sexual homicide: An integrative review", *Aggression and Violent Behavior*, 5(1): 1-22.
- Oliver, C. J., Beech, A. R., Fisher, D., & Beckett, R. (2007). "A comparison of rapists and sexual murderers on demographic and selected psychometric measures", *International Journal of Offender Therapy and Comparative Criminology*, 51(3): 298-312.
- Roberts, J. V., & Grossman, M. G. (1993). "Sexual homicide in Canada: A descriptive analysis" *Annals of Sex Research*, 6(1): 5-25.
- Schlesinger, L. B. (2007). "Sexual homicide: Differentiating catathymic and compulsive murders", *Aggression and Violent Behavior*, 12(2): 242-256.
- Sea, J., & Beauregard, E. (2021). "An analysis of crime scene behavior in Korean homicide", *Journal of Interpersonal Violence*, 36(1-2): 184-207.
- Simundic, A. M. (2012). "Diagnostic Accuracy—Part 1: Basic Concepts Sensitivity and Specificity, ROC Analysis, STARD Statement", *Point of care*, 11(1): 6-8.

Characteristics of Sexual Homicide: Based on Random Forest Analysis

Hyeokjun Kwon* · Jonghan Sea**

Current study is conducted to explore characteristic of sexual homicide in Korea, and to present model which shows best explanation about sexual homicide. Current study uses random forest analysis that is a one of method by machine learning technic to solve methodological problem of preceding studies, and logistic regression to present predicting direction of variables. Current study inputs variables related with offender, victim, and crime scene behaviors. Result in study, our model has moderate level of predicting indicator somewhat, this predictor will be greater by adding cases and information. Relationship with offender and victim(non relationship) is the best predicting variable, and the lowest is location of body discover. Result in regression, offender's sex(men) and marriage status(single) predict sexual homicide positively, and education level predicts sexual homicide negatively. When offender did not have relationship with victim, overkill predicts sexual homicide negatively. Current study has some implications in terms of exploring sexual homicide model to help crime investigation and offender profiling. Also we expect that current study can expand study of sexual homicide in Korea. Current study has academical implication in terms of application for machine learning technic to analysis of sexual homicide.

❖ Key words: Sexual homicide, Random Forest, Machine Learning, Homicide, Crime Prediction

투고일 : 2월 27일 / 심사일 : 3월 23일 / 게재확정일 : 3월 23일
--

* Researcher, Department of Psychology, Yeungnam University

** Professor, Department of Psychology, Yeungnam University