논문 20-45-07-11 The Journal of Korean Institute of Communications and Information Sciences '20-07 Vol.45 No.07
https://doi.org/10.7840/kics.2020.45.7.1210

인공지능 기반 건물화재 예측모델 연구

고 경 석◆, 양 재 경°, 황 동 현*, 고 효 석*, 가 칠 오**, 조 주 필***

Building Fire Prediction Model Study Using AI

Kyeongseok Ko◆, Jaekyung Yang°, Donghyun Hwang*, Hyoseok Ko*, Chillo Ga**, Juphil Cho***

요 약

우리나라는 지난 10년간의 화재발생 건수는 매년 약 4만 건 내외로 발생하고 있으며 점진적으로 감소 추세에 있다. 하지만, 화재규모의 대형화로 인해 화재 건수보다 중요한 인적, 물적 피해는 점점 확대되고 있다. 본 연구에서는 화재사고로 인한 재산피해액, 인명피해 건수가 대형화되고 있는 추세에 효과적으로 대응할 수 있도록 인공지능 기술을 활용하여 데이터 기반의 화재 예측 모델을 개발하고자 하였다. 이를 위해 국토교통부를 중심으로 개방되고 있는 공공데이터를 활용하여 건물 단위로 화재관련 변수를 융합하고, MLP(Multi-Layer Perceptron)의 깊은 신경망 모델을 활용하여 화재예측 모델을 개발하였다. 개발된 모델은 융합데이터셋의 랜덤샘플링한 6만개에 대하여 10-겹 교차검증(10-fold cross validation)을 통해 모델 검증결과 87.1%의 비교적 높은 정확도를 보였다. 이러한 예측 모델의 결과는 건물화재 안전 점검 시, 건물별 화재위험등급을 고려하여 점검 우선순위, 점검 주기를 관리하는 등 화재 예방활동에 활용될 수 있을 것이다.

키워드 : 화재예측, 딥러닝, 인공지능, 텐서플로, 공간정보
Key Words : Fire Prediction, Deep Learning, AI, Tensorflow, Spatial Data

ABSTRACT

In Korea, the number of fires has been around 40,000 per year over the past decade, and is on a gradual decline. However, human and property damage, which is more important than the number of fires, is increasing due to the large scale of fire. In this study, we wanted to develop a data-based fire prediction model using artificial intelligence technology to effectively respond to the growing trend of property damage and human casualties caused by fire accidents. To this end, fire-related variables were fused on a building-by-building basis by utilizing public data being opened to the Ministry of Land, Infrastructure and Transport. Fire prediction model was developed using deep neural network model of the Multi-Layer Perceptron(MLP). The developed model showed relatively high accuracy of 87.1% as a result of the model verification through 10-fold cross validation for 60,000 random sampled units. The result of this predictive model could be used for fire prevention activities, such as management of inspection priority and inspection cycle, considering the fire risk rating of each building during safety inspection of building fires.

※ 본 연구는 '서울시 산학연 협력사업(IC190013)'으로 수행되었습니다.
◆ First Author : UPDATER Corp., Jeonbuk National University, tonyk7845@gmail.com, 박사수료, 정회원
° Corresponding Author : Jeonbuk National University, jkyang@jbnu.ac.kr, 교수, 정회원
* UPDATER Corp.
** Jeonbuk National University, beyondcrow@jbnu.ac.kr, 연구교수, 정회원
*** Kunsan National University, stefano@kunsan.ac.kr, 교수, 정회원
논문번호 : 202004-098-C-RE, Received April 28, 2020; Revised June 18, 2020; Accepted June 22, 2020

Ⅰ. 서 론

우리나라의 화재 발생은 매년 약 4만 건 내외로 2013년까지 조금씩 감소하다가 이후부터는 증가하는 추세를 보이고 있다. 또한 10년간의 평균 화재 건수는 43,374건으로 집계되고 있다. 이러한 집계는 2007년 국가화재정보시스템 구축으로 국가화재분류체계가 만들어지고, 이를 통해 화재 통계가 자동 취합된 결과이다. 객관적 데이터의 확보를 바탕으로 2010년 소방방재청은 '화재와의 전쟁'을 선포하고, 2013년 '화재피해 저감정책'으로 화재 건수를 감소하였으나 다시 해마다 조금씩 증가함을 볼 수 있다[1].

그러나 화재 건수보다 중요한 물적, 인적 피해는 점점 확대되고 있다. 10년간의 평균 인명피해는 2,169명(사망 315명, 부상 1,854명)이고, 재산피해는 382,467백만 원으로 파악되었다. 구체적으로 연도별로 인적피해 관련 데이터를 살펴보면 2016년 2,024명, 2017년 2,197명, 2018년 2,594명으로 증가세가 높고, 물적 재산 피해는 2016년 4,206억 원에서 2017년 5,069억 원, 2018년 5597억 원으로 확대되고 있다[1].

이와 같이 화재발생 건수에 비하여 날로 확대되고 있는 물적/인적 피해를 줄이고, 예방하기 위해서는 화재 위험도를 예측할 수 있는 모델이 필요하다. 과거 통계적 데이터를 바탕으로 새로운 ICT 기술을 활용하여 위험한 건물을 파악하고, 사전에 이에 대한 대비를 할 필요가 있다.

과거 화재 이력에 대한 행정구역별 통계 데이터만으로는 예방 대책을 세우는데 한계가 있다. 이를 해결하기 위해서는 실제 화재가 발생하는 건물 단위의 위험도를 예측하고, 관계 기관에서 활용할 수 있도록 시스템화가 필요하다. 즉, 활용 가능한 예측 정확도가 확보되고 건물 단위 화재위험도 모델 개발이 우선 되어야 한다.

특히, 주요 화재 요인 중 23%를 차지하는 전기적 요인은 수용가(전기사용 고객 정보)를 중심으로 전기점검 데이터를 확보하고 있으며, 해당 수용가의 주소를 기반으로 해당 건물의 속성 정보가 있어 전기적 화재 사고를 기준으로 모델 개발이 가능할 수 있다[2].

기존 연구를 살펴보면, 전문가 집단의 AHP분석을 활용하여 화재위험지수를 작성해왔고, 데이터기반의 분석은 시도-시군구별로 화재위험지수를 산정해오고 있다. 이러한 분석결과는 실제 정책이나 화재안전점검 우선순위 선정 등 예방활동에 활용하기에 어려움이 큰 실정이다.

국외사례를 살펴보면, 미국 아틀란타 소방청(Atlanta Fire Rescue Department, AFRD)에서 구축한 Firebird가 대표적이다. 건물 단위로 화재위험도를 예측하여 아틀란타시의 제한된 시간과 예산으로 효율적인 화재 발생 위험도가 높은 건물의 점검 우선순위를 결정하는 것이 주목적이다. 하지만 랜덤포레스트(Random Forest) 기반으로 개발된 예측 모델은 약 71%의 정확도를 보였으나 약 5,000개의 건물을 대상으로 하였으며, 건물에 관한 속성도 면적, 층수, 필지 등 제한적인 변수만을 활용한 한계가 존재한다[3]. 건물과 관련된 속성은 노후 년 수, 용적률, 건폐율 등 제한적으로 활용함으로써 딥러닝과 같이 가능한 모든 변수를 활용하고자 하는 본 연구와는 차이가 있다.

따라서 본 연구에서는 국토교통부 등에서 관리되고 있는 건물 데이터를 중심으로 화재 관련 융합데이터셋을 생성하고 건물단위로 인공지능 기술을 적용하여 화재 발생 위험도를 예측할 수 있는 모델을 개발하였다. 이러한 결과는 소방청, 지자체 등 관계 기관에서 건물단위 화재 위험 등급을 식별하고 예방활동을 수행할 수 있도록 시스템을 구축하는 기반으로 활용될 수 있다.

Ⅱ. 연구방법

본 연구의 공간적 범위는 전국 허가를 득한 건축물 전체를 대상으로 하였으며, 전국 건물화재 데이터와 국토부 건축물 대장 정보 등 건물속성 중심으로 융합한 데이터이다.

시간적 범위는 2018년도 화재건 수 4만여 건을 대상을 목표 값으로 구성하고, 건축물 정보의 융합, 에너지 정보 등 모든 데이터 기준을 2018년으로 하였다. 이는 실제 화재 데이터를 기반으로 학습 데이터의 시간적 오차를 발생하지 않기 위함이다.

내용적 범위는 소방청의 화재 데이터 및 국토부의 공공데이터의 건축물 관련 데이터 및 행정안전부의 도로명 주소의 데이터 등을 활용하여 데이터의 융합 방법론을 연구하였고, 실제 화재 데이터와 결합한 학습데이터를 구축하고 딥러닝 기법을 활용하여 예측모델을 개발하였다. 예측모델을 활용한 화재사고위험도를 도출하여 유관기관 등에게 정보제공을 위한 서비스 개발을 수행하였다.

본 연구는 전국 600만여 건물에 대한 화재사고에 대한 학습모델을 개발하는데 있는데, 건물에 비해 화재건수가 당연히 작을 수밖에 없다. 이러한 불균형데이터 처리에 대한 연구도 함께 수행하였다. 예측모델링은 MLP(Multi-Layer Perceptron)의 깊은 신경망 모

델을 활용하였고, 모델검증을 위하여 10-겹 교차검증(10-fold cross validation)을 활용하여 모델을 평가하였다. 예측결과를 사용자가 인지하기 쉽도록 위험, 경계, 주의, 관심, 안전 5가지 등급으로 화재위험도를 등급화하기 위하여 K-평균 클러스터링(K-means clustering)을 수행하였다.

[그림 1]은 본 연구의 전체적인 프로세스를 나타낸다.

데이터 수집/전처리 단계에서는 수집된 데이터에서 결측치 등을 제거, 치환을 통해서 전처리하고 건물단위로 결합(join)을 통해서 융합데이터셋을 구축하였다. 데이터분석 단계에서는 주요한 필드를 대상으로 그래프 분석을 통해 탐색적 분석을 수행하고, 상관분석, 통계적 분석을 수행하였다. 예측모델개발 단계에서는 융합셋에서 화재사고 기준으로 1:4비율로 언더샘플링(Under Sampling)을 수행한 후 텐서플로우(Tensorflow)의 학습데이터로 형태로 변환하기 위하여 정규화(Normalization), 원핫인코딩(One-hot Encoding) 등을 수행했다. 이어서 순차(Sequential) 모델을 활용하여 모델을 개발하였다. K-겹 교차검증(K-flold cross validation) 기법을 통해 모델을 선택하고 선택된 모델을 활용하여 화재위험도를 도출하였다. 위험등급도출 단계에서는 클러스터링 기법을 활용하여 위험, 관심, 주의, 경계, 위험 5등급으로 등급화를 수행하였다.

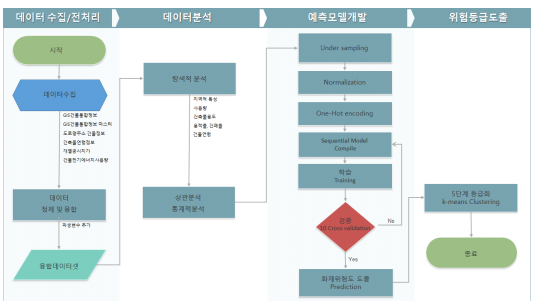

그림 1. 연구과정
Fig. 1. Research Process

Ⅲ. 예측모델링

3.1 데이터수집

건물단위로 데이터를 융합하여 화재 예측위험도를 도출하는 것이 목적이기 때문에 국가공간정보포털에 있는 건물속성정보와 건물공간정보(shp), 행정안전부 도로명 건물데이터 등 다양한 독립변수를 수집하였다.

표 1. 데이터 목록
Table 1. Data List

Data	Data field	Provider
GIS building integrated info	UFID, Building Floors, Use, Building Structure, …	National spatial info portal
Individual Land Prices	Land area, Price, PNU, …	
Building age info	Building age, area, height, …	
GIS building info master	UFID, Building management number	
Road Name Address Building Info	Building management number, city name	Road name address info system
Building electricity usage info	electricity consumption, address, ...	Architecture Administration System
Administrative district	administrative borders …	National spatial info portal
Fire accident info	address, building management number	

국가공간정보포털은 산재되어 있는 공간정보를 활용하기 쉽도록 국가, 공공, 민간에서 생산한 공간정보를 한곳에서 활용할 수 있도록 구축한 서비스이다[4].

국가공간정보포털, 도로명주소안내시스템, 건축행정시스템(세움터)에서 [표 1]과 같이 8개의 데이터셋을 수집하였다.

3.2 데이터 전처리 및 융합

수집된 데이터에서 불필요한 속성을 제거 등 전처리를 수행했고, 건물고유번호를 중심으로 건물단위로 데이터를 융합하여 최종 융합셋을 구축하였다.

공간객체등록번호(UFID)는 국토부에서 사용하는 고유번호를 말하고 공간정보에 표시된 건물 등 시설물에 식별번호를 부여한 것이다. 건물 등 주요시설물 정보를 데이터베이스화해 지리정보시스템에서 활용하기 위해 개별시설물마다 부여한 국가표준 식별번호다. 문자와 숫자를 조합해 17자리로 구성된다.

건물고유번호(BD_MGT_SN)는 행안부에서 사용하는 건물단위를 구분하기 위한 고유번호이다. 도로명주소 식별용 조합키로 도로명코드(12)+읍면동일련번호(2)+지하여부(1)+건물본번(5)+건물부번(5)로 구성되어 있고 총 25자리로 구성된다.

타기관의 데이터를 건물단위로 융합하기 위하여 매

핑테이블 역할을 하는 GIS건물통합마스터 데이터를 활용했다.

융합과정을 살펴보면, GIS건물통합정보마스터(27개 속성, 6,268,145개 인스턴스)와 도로명주소건물정보(31개 속성, 10,740,114개 인스턴스) 데이터 건물고유번호(BD_MGT_SN)로 융합하면 50개 속성, 6,034,472개 인스턴스를 가진 1차 융합셋 도출하였다. 1차 융합셋(50개 속성, 6,034,472개 인스턴스)과 GIS건물통합정보(23개 속성, 13,885,976개 인스턴스)를 UFID로 융합하여 72개 속성, 5,963,167개 인스턴스를 가진 2차 융합셋을 도출하였다. 2차 융합셋에서 중복 및 불필요한 속성(19개 속성) 삭제, 46개 속성 이름변경 후 53개 속성으로 정리하였다. 2차 융합셋에 지역구분(특별시, 광역시, 일반시, 일반군)으로 구분할 수 있는 파생변수를 추가하였고, 건축물 용도코드 매핑테이블을 활용하여 건물의 용도명 속성을 추가하였다. 이어서 건물연령정보(20개 속성, 5,763,745개 인스턴스), 개별공시지가(8개 속성, 33,530,997개 인스턴스), 전기에너지사용량정보(53개 속성, 1,371,074개 인스턴스)를 융합하여 융합데이터셋을 도출하였다.

이 융합데이터셋에 종속변수인 일반화재사고(9개 속성, 38,397개 인스턴스 - 2018년 전체화재사고)데이터 건물고유번호로 융합하여 화재유무 속성을 추가하였다. 도로명주소건물 공간정보와 융합데이터셋을 건물고유번호로 결합을 통해 [그림 2]와 같이 공간정보기반의 건물속성정보를 포함하는 데이터셋을 구축하였다.

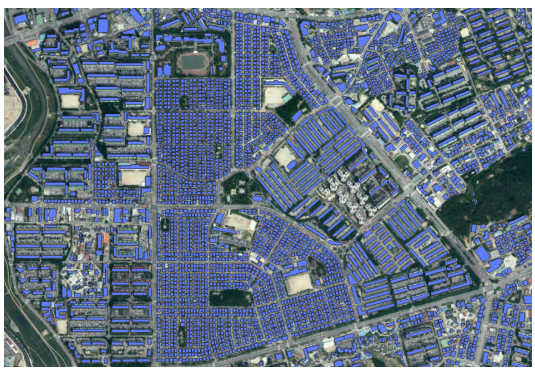

그림 2. 지오코딩(공간화)
Fig. 2. Geocoding

3.3 데이터 분석

지역적 구분으로 건물 수 기준으로 화재사고의 수를 비교하면 일반시, 특별시, 광역시, 일반군 순이나, 이를 해당 구분의 건물수로 나누어 계산하면 특별시, 광역시, 일반시, 일반군 순으로 화재가 많이 발생하는 것을 볼 수 있다. 화재발생의 특징이 지역적인 차이가 있다는 가설에 의해서 특별시, 광역시, 일반시, 일반군으로 구분하는 속성을 만들었다.

[표 2]와 같이 건물데이터의 수는 일반시, 일반군, 광역시, 특별시 순이지만 에너지사용량 평균값은 역순으로 특별시, 광역시, 일반시, 일반군으로 많았다. 에너지 사용량이 높으면 화재위험이 높아진다는 것을 볼 수 있다.

단독주택 60%, 공동주택 8%, 제2종 근린생활시설 8%, 제1종 근린생활시설 7%로 가장 많고 일반시, 일반군은 공동주택의 비율이 줄어드는 것을 볼 수 있다. 업무시설의 화재율이 1.6%로 가장 높고, 공동주택 0.7%, 공장 0.4%순으로 나타났다.

용적률은 대지면적에 대한 연면적의 비율로, 활용된 데이터셋에서 화재가 일어난 건물들의 평균 용적률이 높은 특성을 나타내었다. 따라서 건물들의 용적률이 높을수록 건물밀도가 높아져 화재에 취약할 수 있다는 가정 하에 설명변수로 활용하였다.

건물건축일자로부터 건물연령을 도출하여 그 분포를 [그림 3]과 같이 확인하였다. 15년 이상 40년 이하 건물이 데이터의 86%를 차지하고 있었다.

건물연령대별 화재비율을 계산해보면, [표 3]과 같이 15년 이상 30년 이하 건물이 가장 높은 비율로 화

표 2. 지역별 건물수 대비 화재건수
Table 2. Usage data list

Regional classification	Number of fires	Number of buildings	Ratio
Metropolitan (Teugbyeol-si)	3,789	529,789	0.71%
City (gwangyeog-si)	3,764	1,031,499	0.36%
Town(Si)	4,661	3,183,090	0.14%
Country(Gun)	14	1,218,754	0.001%

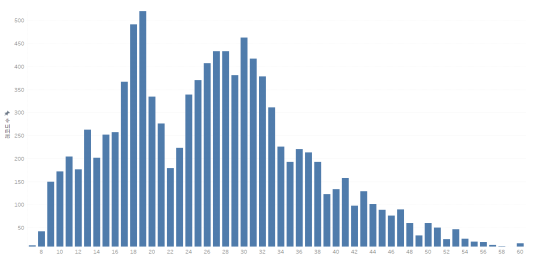

그림 3. 건물연령 시각화
Fig. 3. Building age visualization

표 3. 건물연령대별 화재비율
Table 3. Usage data list

Building age	Number of fires	Number of buildings	ratio
15 years or less	1,480	420,392	0.35%
15 to 30 years	5,537	1,416,356	0.39%
30 to 45 years	2,989	974,734	0.30%
45 to 60 years	554	302,135	0.18%

재사고가 발생했고 15년 이하 건물과 30년 이상 45년 이하 건물과 비교해보면 큰 차이는 보이지 않았다.

3.4 예측모델 개발

모델개발을 위한 방법론을 크게 샘플링기법 선택, 학습 데이터셋 구축을 위한 데이터변환, 모델링, 하이퍼파라미터 최적화 순으로 수행하였다.

언더샘플링의 비율에 따른 모델성능을 비교 분석하기 위하여 10:90, 20:80, 30:70, 50:50 으로 샘플링 후 모델평가를 수행하고 AUC(Area Under ROC Curve)를 비교하였다.

ROC(Receiver Operator Characteristic)는 모델의 성능이 기준선을 넘었는지 확인하기 위해 시각적으로 표현한 그래프다. ROC 그래프는 FPR을 X축으로, TPR(Sensitivity) 을 Y축으로 정의 한다. 각각의 예측은 ROC 공간에 점으로 표현될 수 있다. 완벽한 모델의 경우 모든 데이터 포인트에 대해 TPR은 1이 되고 FPR은 0이 된다. 평균 모델이나 기본선 모델은 (0,0)에서 (1,1)까지 대각선으로 표현할 수 있다. 이는 TPR 값과 FPR 값이 0.5가 되는 것을 나타낸다. 만약 우리 모델의 ROC 곡선이 기본 대각선보다 위에 있다면, 이 모델의 성능이 기본선보다 좋음을 나타낸다. Area Under Curve는 평가 모델의 ROC곡선 아래 영역의 넓이를 의미한다. AUC 값은 모델에서 항목을 임의로 추출했을 때 긍정 항목이 부정 항목보다 더 선택될 확률을 나타낸다. 따라서 높은 AUC가 더 좋은 모델이다[5].

화재데이터의 수는 전체 건물수에 비해 너무 작기 때문에(Imbalanced data) 오버샘플링을 진행 할 경우 데이터를 너무 많이 복제해야하기 때문에 오버피팅의 위험이 있으므로, 전체 융합셋(5,965,126)에서 화재사고 중심으로 1:4의 비율로 언더샘플링하여 학습데이터 구축하였다.

샘플링에 따른 모델평가 성능척도를 ROC Curve와 AUC를 활용했다.

언더샘플링 비율에 따른 데이터를 활용하여 모델을 평가한 결과 [그림 4, 5]와 같이 20:80의 비율로 언더

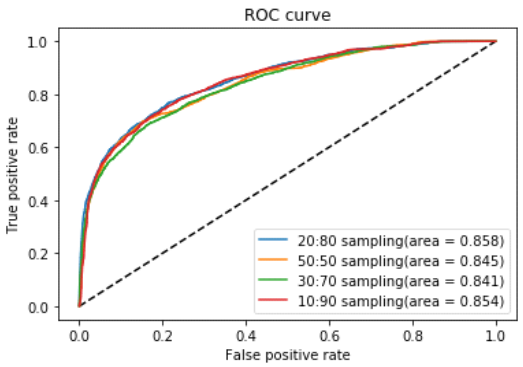

그림 4. Sampling비율에 따른 ROC curve
Fig. 4. ROC curve according to the sampling rate

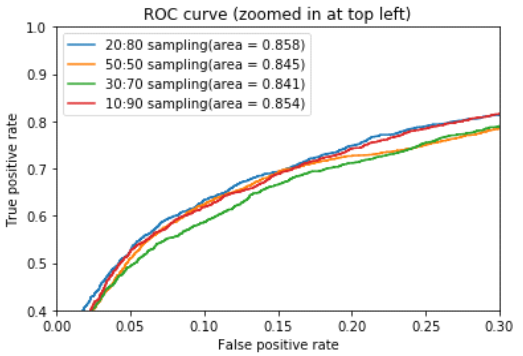

그림 5. Sampling비율에 따른 ROC curve(확대)
Fig. 5. ROC curve according to the sampling rate (enlargement)

샘플링한 모델의 AUC가 0.858로 가장 좋게 나타났다.

학습을 위해서 전기사용량, 개별공시지가, 면적 등 연속형 변수에 대해서 0과 1사이의 값으로 최대-최소 스케일러(Min-Max Scaler)를 활용한 정규화를 수행하였다. 명목형 변수에 대해서는 원핫인코딩을 수행하여야 한다. '지목코드', '건축물용도명', '건축물구조명', '지역구분', '업종' 5개 필드를 원핫인코딩을 수행하였다.

최대-최소 스케일러는 데이터를 정규화 하는 가장 일반적인 방법으로 모든 변수가 0과1사이의 값으로 스케일을 조정하는 방법이고, 모든 변수가 스케일이 동일하지만 이상치를 잘 처리하지 못하는 단점이 있다[6].

'건물고유번호', '사용_량(KWh)_Mean', '지목코드', '개별공시지가', '대지면적', '건축물용도명', '건축물구조명', '건축물면적', '높이_건통', '건폐율_건통', '용적율_건통', '지상층수', '지하층수', '지역구분',

'건물연면적', '건물연령', '업종' 17개의 설명변수와 '화재유무'를 종속변수로 설정하여 Keras의 순차모델을 만들었다. 총 12개의 계층(layer) 생성하였고, 층이 깊어져 과적합 방지를 위해 Dropout 레이어를 추가하였다.

Dropout은 신경망의 깊이가 깊어질수록 과적합이 일어날 확률이 증가하기 때문에 일부의 뉴런을 확률적으로 생략시킴으로써 과적합을 방지하기 위해 사용된다. [그림 6]은 Dropout을 도식화한 그림이다.

순차모델을 활용하여 아래와 같이 모델을 생성하였다. 입력값은 99개 필드로 구성되어 있고, dropout 레이어, 활성함수(activation function)를 추가하였다.

[그림 7]과 같은 인공신경망을 구축하여 4,971개의 파라미터를 학습하였다. 활성함수로는 Relu를 사용하였고 마지막 레이어는 시그모이드(sigmoid)를 사용하였다. 손실함수(loss function)는 이진 크로스엔트로피(binary crossentropy)로 설정하고, 최적화는 Adam을 활용하였다. 배치사이즈(batch size)는 40, 학습 횟수(epoch)는 40,검증은 훈련데이터셋의 10% 활용하여 학습을 수행하였다.

하이퍼파라미터 최적화를 위하여 [표 4]와 같은 검토리스트를 활용하여 그리드서치(Grid Search)를 수행하였다. 그리드서치는 리스트로 지정된 하이퍼파라미터 값 전체를 검토하여 모든 조합에 대해 모델 성능을 평가하여 최적의 조합을 찾는 기법을 말한다.

그 결과 batch_size = 40, epochs = 60, learning_rate = 0.001, dropout_rate = 0.0, optimizer = 'Adam', activation = 'relu'의 하이퍼파라미터 값이 가장 좋은 성능을 나타내는 것으로 나타났다.

학습 횟수에 따른 검증 손실(validation loss), 검증 정확도(validation accuracy) 변화량 그래프로 그려보면, [그림 8, 9]와 같이 학습이 진행 될수록 검증 손실는 대체적으로 낮아지고, 검증 정확도는 점차 높아지는 것을 알 수 있다.

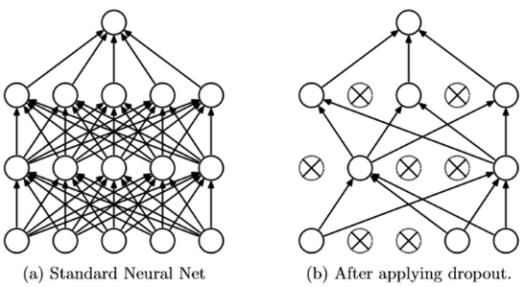

그림 6. dropout 개념도
Fig. 6. dropout concept diagram

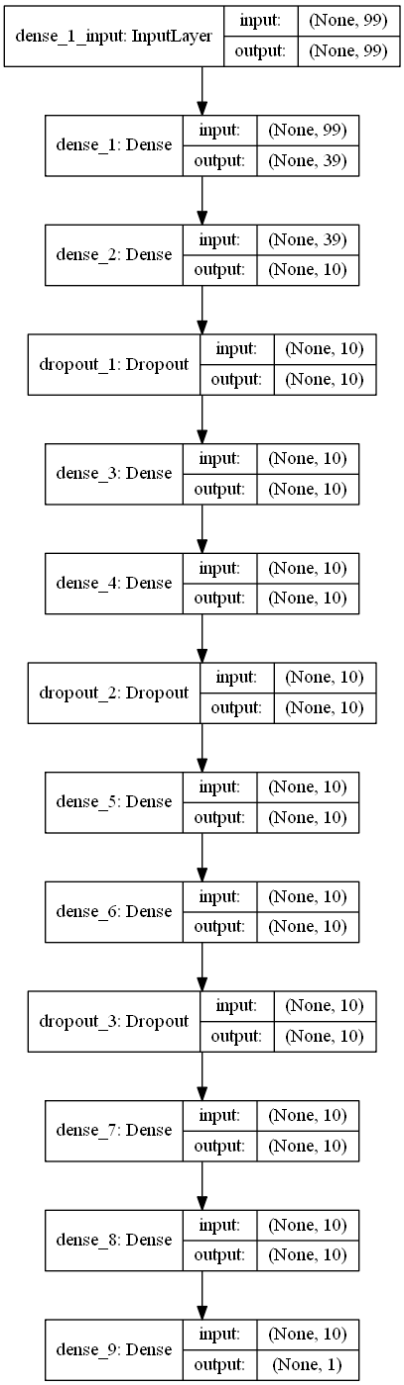

그림 7. 인공신경망 구조
Fig. 7. Artificial neural network structure

테스트셋을 활용하여 모델 테스트 결과 최종 예측 정확도는 86.91%로 나왔다.

표 4. grid search 검토 리스트
Table 4. grid search check list

Hyperparameter	range
batch_size	10, 20, 40, 60, 80
epochs	20, 40, 60, 80, 100
learning_rate	0.0001, 0.001, 0.01, 0.1
dropout_rate	0.0, 0.05, 0.1, 0.15, 0.2, 0.3, 0.4, 0.6
optimizer	'SGD', 'RMSprop', 'Adagrad', 'Adadelta', 'Adam', 'Nadam'
activation	'softmax', 'softplus', 'softsign', 'relu', 'tanh', 'sigmoid', 'linear'

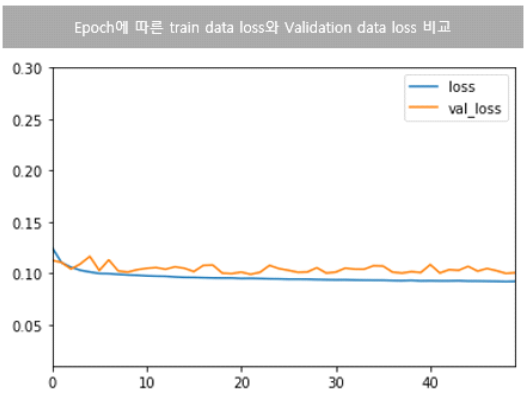

그림 8. epoch에 따른 loss 변화 그래프
Fig. 8. Graph of change in loss according to epoch

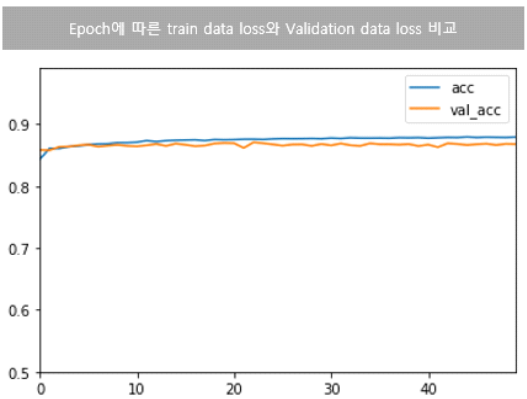

그림 9. epoch에 따른 accuracy 변화 그래프
Fig. 9. Graph of change in accuracy according to epoch

3.5 검증

k-겹 교차검증(k-fold cross validation)은 전체 데이터의 일부를 검증 데이터셋으로 사용하는 모델은 테스트셋을 어떻게 잡느냐에 따라 성능이 달라지기 때문에 우연의 효과로 모델 평가 지표에 편향이 생기는 문제가 있지만 k-겹 교차검증은 모든 데이터가 최소 한번은 테스트 셋으로 쓰이기 때문에 그러한 문제를 보완해준다[7]. [그림 10]은 k-겹 교차검증의 개념을 도식화한 그림이다. 훈련, 테스트셋을 통해 진행하는 일반적인 학습법에 비해 시간 소요가 크지만 총 개수가 적은 데이터셋에 대하여 성능 평가의 신뢰성을 준다.

10-겹 교차 검증기법을 활용하여 만들어진 모델의 검증을 수행하였다. 10-겹 교차 검증 학습 단계별 검증 결과를 보면, [그림 11]과 같이 검증결과 정확도는 87.1%로 안정적인 결과가 도출되었다.

클러스터링 간에 비교하여 적절한 클러스터 개수를 구하기 위해서 Elbow method를 활용하여 클러스터개수를 선정하였다. K-평균 클러스터링은 클러스터내의 SSE(Sum of Squared Errors)의 값이 최소가 되도록 클러스터 중심을 결정을 하는 방법이다. 클러스터 개수가 늘어날수록 SSE값이 작아지는데, SSE 값이 점차 줄어들다가 줄어드는 비율이 급격하게 작아지는 지점(elbow)을 도출하여 최적의 클러스터 개수로 지정하였다[8].

K-평균 클러스터링의 cost function은 아래와 같다.

$$Minimize\, C_1,\ldots,C_k \sum_{k=1}^{K} W(C_k) \tag{1}$$

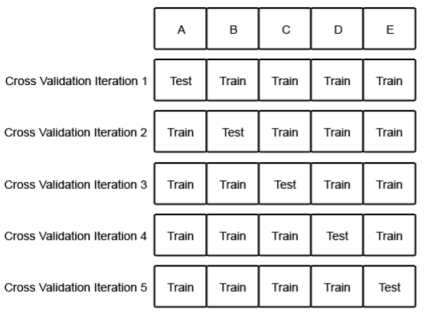

그림 10. k-fold cross validation 개념도
Fig. 10. Conceptual diagram of k-fold cross validation

```
cv = KFold(10, shuffle=True, random_state=0)

results

array([0.87337028, 0.88300782, 0.8666884 , 0.86408083, 0.8714146 ,
       0.87353325, 0.87239244, 0.87157758, 0.86829666, 0.86422168])

np.mean(results)

0.8707583631070716
```

그림 11. 검증결과
Fig. 11. Verification Result

예측결과에 elbow method 적용결과 [그림 12]의 결과를 보였으며, 이를 토대로 최적의 클러스터 개수를 5로 적용하여 [표 5]와 같은 화재위험등급을 도출하였다.

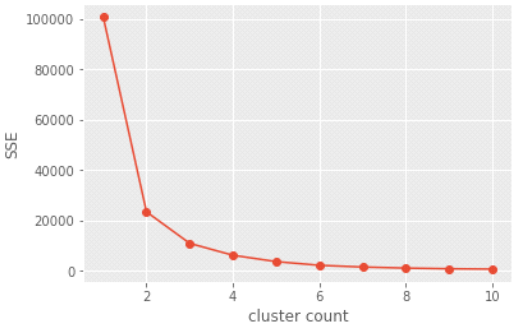

그림 12. Elbow method 수행결과
Fig. 12. Results of Elbow method

표 5. 위험등급에 따른 집단 수
Table 5. Number of groups according to risk level

Division	Cluster name	Count	Ratio
A grade (Safety)	group-1	4,496,710	78.76%
B grade (Attention)	group-5	712,069	11.94%
C grade (Caution)	group-3	305,470	5.12%
D grade (Alert)	group-2	209,532	3.51%
E grade (Danger)	group-4	39,351	0.66%
Total	5,963,132	5,963,132	100%

Ⅳ. 결 론

본 연구는 다양한 요인들의 복합적 작용으로 발생하는 화재사고에 대응하기 위해 인공지능 기법을 적용하여 보다 객관화하고 계량화할 수 있는 예측 기법을 연구하였다.

우선, 건물화재사고 예측을 위해 국토부의 건물 중심 속성데이터, GIS건물통합정보, 건물연령정보, 개별공시지가정보, 건물에너지사용량정보 및 도로명주소 건물정보를 공간정보기술을 활용하여 건물단위 분석이 가능하도록 융합셋으로 구축하였다.

구축된 융합셋을 활용하여 건물단위 화재위험 예측 모델을 인공지능 기술을 활용하였고, 그 중에서 MLP의 깊은 신경망 모델을 통해서 학습하였다. Grid Search 기법을 통해 하이퍼파라미터 최적화 과정을 거쳐 건물화재 예측 학습모델을 개발하였다. 10-겹 교차검증의 모델 검증과정을 거쳐 도출된 결과 87.1%의 정확도를 확보하였다. 예측결과를 사용자가 인지하기 쉽도록 위험, 경계, 주의, 관심, 안전 5가지 등급으로 화재위험도를 등급화하기 위하여 K-평균 클러스터링을 수행하였다.

또한, GIS 기술을 활용한 건물화재 위험도지도 제공을 위하여 분석 및 예측한 결과를 화재위험도 5등급화를 통한 분석결과를 지도기반 시각화를 수행하였다.

본 연구에서 구축한 건물단위 융합데이터셋은 다양한 스마트시티 활용 모델을 개발하는데 있어 기반 데이터로 활용될 수 있을 것으로 판단되며 또한, 공간정보와 인공지능 기술을 융합하여 보다 과학적으로 화재 예측을 할 수 있는 알고리즘을 제시한데 의의가 있다고 할 수 있다.

아울러 본 연구를 토대로 유관 기관들이 실시간으로 화재 위험 정보를 확인할 수 있도록 시스템을 구축하여 제공한다면 매우 효율적인 공공 서비스가 될 수 있을 것으로 판단된다.

References

[1] Information Statistics Officer, "*Fire Statistics Yearbook*," National Fire Agency, 2018.

[2] K. Ko, "Electrical fire prediction model study using machine learning," *J. KIIECT*, vol. 11, no. 6, pp. 703-710, 2018.

[3] M. Madaio, S. T. Chen, O. L. Haimson, W. Zhang, X. Cheng, M. Hinds-Aldrich, D. H. Chau, and B. Dilkina, "Firebird: Predicting fire risk and prioritizing fire inspections in Atlanta," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, pp. 185-194, San Francisco, CA, USA, Aug. 2016.

[4] *Spatial Information Portal*, nsdi.go.kr. 2020.

[5] J. Kim and H. Park, "Imbalanced data analysis using sampling methods," Inha Univ., pp. 6-7, Feb. 2012.

[6] S. Gopal Krishna Patro and Kishore Kumar Sahu, "*Normalization: A preprocessing stage*,"

ArXiv abs/1503.06462, 2015.

[7] S. Russell and P. Norvig, "*Artificial Intelligence: A Modern Approach*," Pearson Education, 3rd Ed., p. 181, 2013.

[8] P. Bholowalia and A. Kumar, "EBK-means: A clustering technique based on elbow method and k-means in WSN," *Int. J. Comput. Appl.*, vol. 105, no. 9, Nov. 2014.

고 경 석 (Kyeongseok Ko)

2006년 6월 : 서울교통공사 정보화기획단

2011년 5월 : 한국국토정보공사 공간정보사업실

2015년 3월~현재 : 전북대학교 산업시스템공학과 박사수료

<관심분야> 머신러닝, 인공지능, 공간정보

[ORCID:0000-0001-7598-185X]

양 재 경 (Jaekyung Yang)

2003년 12월 : Iowa State University 산업공학 박사, 전산학 부전공

2004년~현재 : 전북대 산업시스템공학과 교수

<관심분야> 기계학습, 정보시스템

[ORCID:0000-0002-4904-1351]

황 동 현 (Donghyun Hwang)

2019년 8월 : 전북대학교 경영학과 경영학박사

2013년~2018년 : 한국국토정보공사 팀장

2018년~현재 : 주식회사 업데이터 대표이사

<관심분야> 빅데이터, 인공지능, 예측서비스

고 효 석 (Hyoseok Ko)

2016년 2월 : 경남대학교 학사 졸업

2018년 12월~현재 : 업데이터 연구원 재직

<관심분야> 머신러닝, 딥러닝, 빅데이터

가 칠 오 (Chillo Ga)

2013년 8월 : 서울대학교 건설환경공학부 졸업

2019년 10월 : 한국국토정보공사 기획조정실 차장

2019년 11월~현재 : 전북대학교 빅데이터비즈니스연구소 연구교수

<관심분야> 공간정보, 인공지능, 빅데이터

조 주 필 (Juphil Cho)

2001년 2월 : 전북대학교 전자공학과 공학박사

2000년~2007년 : ETRI 이동통신연구단 선임연구원

2011년~2012년 : 미국USF 교환교수

2005년~현재 : 군산대학교 IT융합통신공학과 교수

<관심분야> LTE-A, 5G 이동통신, 차세대 WLAN