발간등록번호 11-B552989-000601-01	**최종보고서(일반과제용)**	보안등급 일반[✔], 보안[ ]

중앙행정기관명	국토교통부		사업명	사업명	국토교통기술촉진연구		
전문기관명	국토교통과학기술진흥원			내역사업명	국토교통기술촉진연구		
공고번호	공고-제05호		총괄연구개발 식별번호		C157011		
			연구개발과제번호		21CTAP-C157011-02		
국가과학기술 표준분류	EE	40%	EE01	30%	EE0108	30%	

총괄연구개발명	국문	교통안전을 위한 상황 딥러닝 기반 이머징 헬스 리스크 마이닝
	영문	Context Deep Learning based Emerging Health Risk Mining for Traffic Safety
연구개발과제명	국문	교통안전을 위한 상황 딥러닝 기반 이머징 헬스 리스크 마이닝
	영문	Context Deep Learning based Emerging Health Risk Mining for Traffic Safety

주관연구기관	기관명	경기대학교 산학협력단	사업자등록번호	135-82-10872
	주소	(16227) 경기도 수원시 영통구 광교산로 154-42, 경기대학교 산학협력단	법인등록번호	135871-0005863

주관연구책임자	성명		정경용	직위	부교수
	연락처	직장전화	031-249-1382	휴대전화	010-6362-4555
		전자우편	kychung@kgu.ac.kr	국가연구자번호	10150345

연구개발기간	전체	2020. 04. 13 - 2021. 12. 31(21개월)
	1년차	2020. 04. 13 - 2020. 12. 31 (9개월)
	2년차	2021. 01. 01 - 2021. 12. 31(12개월)

연구개발비 (단위 : 천원)	정부지원 연구개발비	기관부담 연구개발비		그 외 기관 등의 지원금				합계			연구개발비 외 지원금
				지방자치단체		기타( )					
	현금	현금	현물	현금	현물	현금	현물	현금	현물	합계	
총계	350,000			-	-	-	-	350,000		350,000	-
1년차	130,000			-	-	-	-	130,000		130,000	-
2년차	220,000			-	-	-	-	220,000		220,000	-
				-	-	-	-				-

공동연구기관 등	기관명	책임자	직위	휴대전화	전자우편	비고	
						역할	기관유형
공동연구기관	상지대학교 산학협력단	박찬홍	조교수	010-9286-7263	roypark1984@gmail.com	공동	대학
위탁연구기관							
연구기관 외 기관							

연구개발담당자 실무담당자	성명		백지원	직위	대학원생(박사과정)
	연락처	직장전화	031-249-1382	휴대전화	010-9453-1640
		전자우편	hy1233hh@naver.com	국가연구자번호	11477308

이 최종보고서에 기재된 내용이 사실임을 확인하며, 만약 사실이 아닌 경우 관련 법령 및 규정에 따라 제재 처분 등의 불이익도 감수하겠습니다.

2022년 1월 24일

주관연구책임자 : 정경용 (직인)

주관연구기관의 장 : 경기대학교 산학협력단 윤여강 (직인)

국토교통과학기술진흥원장 귀하

< 요 약 문 >

<table>
<tr><td colspan="3">사업명</td><td colspan="2">국토교통기술촉진연구</td><td>총괄연구개발 식별번호</td><td colspan="2">C157011</td></tr>
<tr><td colspan="3">내역사업명</td><td colspan="2">국토교통기술촉진연구</td><td>연구개발과제번호</td><td colspan="2">21CTAP-C157011-02</td></tr>
<tr><td colspan="3">국가과학기술
표준분류</td><td colspan="2">EE 40% | EE01 30% | EE0108 30%</td><td colspan="3"></td></tr>
<tr><td colspan="3">총괄연구개발명</td><td colspan="5">교통안전을 위한 상황 딥러닝 기반 이머징 헬스 리스크 마이닝</td></tr>
<tr><td colspan="3">연구개발과제명</td><td colspan="5">교통안전을 위한 상황 딥러닝 기반 이머징 헬스 리스크 마이닝</td></tr>
<tr><td colspan="3">전체 연구개발기간</td><td colspan="5">2020. 04. 13 - 2021. 12. 31(1년 9개월)</td></tr>
<tr><td colspan="3">총 연구개발비</td><td colspan="5">총 350,000 천원
(정부지원연구개발비 : 350,000 천원, 기관부담연구개발비 : 천원,
지방자치단체 : 천원, 그 외 지원금 : 천원)</td></tr>
<tr><td colspan="3">연구개발단계</td><td colspan="2">기초[ ] 응용[✔] 개발[ ]
기타(위 3가지에 해당되지 않는 경우)[ ]</td><td>기술성숙도</td><td colspan="2">착수시점 기준(3)
종료시점 목표(4)</td></tr>
<tr><td colspan="3">연구개발과제 유형</td><td colspan="5">지정공모 [✔] 자유공모 [ ]</td></tr>
<tr><td colspan="3">연구개발과제 특성</td><td colspan="5">실용화</td></tr>
<tr><td rowspan="5">연구
개발
목표
및
내용</td><td colspan="2">최종 목표</td><td colspan="5">• 본 연구의 최종 목표는 과학기술적 측면에서 교통기술과 IT 융복합 기술을 통해 교통안전을 위한 상황 딥러닝 기반 이머징 헬스 리스크 탐지 및 마이닝 예측 기술을 개발하여 교통사고 감소를 통한 사회적 비용 절감과 국민 삶의 질 향상에 기여</td></tr>
<tr><td colspan="2">전체 내용</td><td colspan="5">• (주관연구기관) 교통사고 저감을 위한 상황모션 추적 기반의 행동변화 탐지 및 상황 추론, 잠재적 위험에 따른 도로교통 이용자별 교통안전지수, 시공간 결측보정을 통한 LSTM 기반 교통 혼잡도 예측 모델 개발, 빅데이터 기반 교통안전 이머징 리스크 마이닝 예측 기술을 개발
• (공동연구기관) 교통사고 데이터 기반 사고발생 빈도 간 상관 분석을 실시하고, 운전자의 내외부 환경 요인 분석을 통한 사고 위험도 예측 모형을 개발하며, 이를 기반으로 한 사고 위험도 예측 시스템 및 모바일 서비스를 개발</td></tr>
<tr><td rowspan="2">1년차</td><td>목표</td><td colspan="5">교통안전을 위한 상황 딥러닝 모델 및 교통안전지수 개발</td></tr>
<tr><td>내용</td><td colspan="5">• (1차년도) 교통안전 서비스를 위해 다양한 요인분석과 상황 딥러닝 예측 모델을 개발하고 이를 기반으로 교통안전지수 개발에 관한 연구 수행
- 주관연구기관 : 교통안전을 위해 다양한 요인 분석을 통한 교통안전지수 개발에 관한 연구를 진행하며, 교통안전 및 위험관리 분석, 위험예측을 위한 다중회귀 및 상황 딥러닝 모델, 이머징 리스크 상황 예측, 상황 모션 추적 기반의 행동변화 탐지 및 상황 추론, 위험상황 인식 기반 위험패턴 발견 및 교통안전지수 개발에 관한 연구 수행
- 공동연구기관 : 교통사고 데이터 기반 사고 발생 빈도 간 상관 분석을 실시하고, 주관연구기관의 요인분석을 반영한 사고 위험도 예측 모형 개발에 관한 연구 수행</td></tr>
<tr><td rowspan="2">2년차</td><td>목표</td><td colspan="5">교통사고 위험도 예측을 위한 멀티미디어 서비스</td></tr>
<tr><td>내용</td><td colspan="5">• (2차년도) 1차년도에 개발된 교통사고 위험 예측 모델에 대한 고도화 및 검증, 이를 기반으로 한 인터렉티브 휴먼 인터페이스 소프트웨어 개발 및 사고 위험도 예측을 위한 실시간 멀티미디어 서비스 구현에 관한 연구 수행
- 주관연구기관 : 정확한 위험도 예측 서비스 제공을 위해 시공간 결측 보정을 통한 LSTM 기반 잠재적 교통사고 위험도 분석, 데이터 특성에 따른 이상치 및 결측 유형 보정, LSTM 기반 리스크 예측 모델을 통한 임베딩 고도화 및 이머징 리스크 마이닝, 인터렉티브 휴먼 인터페이스 개발에 관한 연구 수행
- 공동연구기관 : 주관연구기관에서 구축한 사고 위험도 예측 모델들과의 유기적 연계를 통해 사고 위험도 예측 실시간 모바일 서비스 개발 및 도심지 교통사고 위험도 예측 시스템의 체계화 및 고도화에 관련된 연구 수행</td></tr>
</table>

연구개발성과

• 교통안전을 위한 상황 딥러닝 기반 이머징 헬스 리스크 마이닝 기술을 개발하여 **SCI급 국제학술지 12건, 국내학술지 게재 4건, SW프로그램저작권 등록 3건, 특허출원 3건** 연구실적물 배출

구분	논문 (SCI)	논문 (비SCI)	특허	보고서 원문	연구시설·장비	기술요약 정보	소프트웨어	화합물	생명자원		신품종	
									생명정보	생물자원	정보	실물
성과	12	4	3				3					

- (1차년도) **SCI급 국제학술지 7건, 국내학술지 게재 2건**

구분	논문 (SCI)	논문 (비SCI)	특허	보고서 원문	연구시설·장비	기술요약 정보	소프트웨어	화합물	생명자원		신품종	
									생명정보	생물자원	정보	실물
성과	7	2										

- (2차년도) **SCI급 국제학술지 5건, 국내학술지 게재 2건, SW프로그램저작권 등록 3건, 특허출원 3건**

구분	논문 (SCI)	논문 (비SCI)	특허	보고서 원문	연구시설·장비	기술요약 정보	소프트웨어	화합물	생명자원		신품종	
									생명정보	생물자원	정보	실물
성과	5	2	3				3					

연구개발성과 활용방안 및 기대효과

• 성과활용계획

- 지능형 도로교통체계에서 위험 상황인식과 이머징 헬스 리스크를 이용한 도로교통 이용자별 교통안전지수를 제공하여 국민 삶의 질 향상
- 교통산업과 IT융합 포럼의 활용 산/학/연 공동 활동을 통해 구체적인 협력 융복합 분야를 도출하고 중점 추진 분야와 분야별 구체화 가능함

• 기대효과

- 교통안전, 인지과학, 정보과학, 상황인식, 헬스 리스크 등의 창의도전 융복합 학문의 접근을 통한 새로운 학문적/기술적 패러다임으로 활용이 가능함
- 도로교통 이용자에 대한 위험 상황인식을 활용하여 위험도 노출에 따른 교통사고 유발 가능성이 있는 잠재적 위험도와 교통안전지수 제공은 모바일 디바이스에서 가능함
- 사용자의 건강상태와 차량 주행 중의 외부 환경의 변화에 따른 상황 정보를 이용하여 컨텍스트 추론 기술 개발, 결과를 이용한 개인화 서비스가 가능함
- 공공기관에서 제공하는 생활기상지수, 보건기상지수 등과 연계하여 교통안전지수의 고도화 정보를 제공하는 노하우 기술이전이 가능함
- 도로교통 이용자에 대한 교통안전지수 모니터링과 사고위험 예측 모델을 사용자 친화적 멀티모달 인터페이스 형태의 모듈 적용이 가능함
- 교통기술과 IT 융복합 기술을 통하여 국민의 실생활과 밀접하게 관련이 있는 교통안전 기술개발로 사회문제 해결 및 국민 삶의 질 향상 기여

연구개발성과의 비공개여부 및 사유	해당없음

연구개발성과의 등록·기탁 건수	논문	특허	보고서 원문	연구시설·장비	기술요약 정보	소프트웨어	표준	생명자원		화합물	신품종	
								생명정보	생물자원		정보	실물
	16	3				3						

연구시설·장비 종합정보시스템 등록 현황	구입기관	연구시설·장비명	규격 (모델명)	수량	구입 연월일	구입가격 (천원)	구입처 (전화)	비고 (설치장소)	ZEUS 등록번호

국문핵심어 (5개 이내)	이머징 리스크	상황 딥러닝	차량 운전	헬스 리스크	교통 안전
영문핵심어 (5개 이내)	Emerging Risk	Context Deep Learning	Vehicle Driving	Health Risk	Traffic Safety

〈 목 차 〉

1. 연구개발과제의 개요

2. 연구개발과제의 수행 과정 및 수행내용

3. 연구개발과제의 수행 결과 및 목표 달성 정도

4. 목표 미달 시 원인분석

5. 연구개발성과 및 관련 분야에 대한 기여 정도

6. 연구개발성과의 관리 및 활용 계획

별첨자료(참고문헌 등)

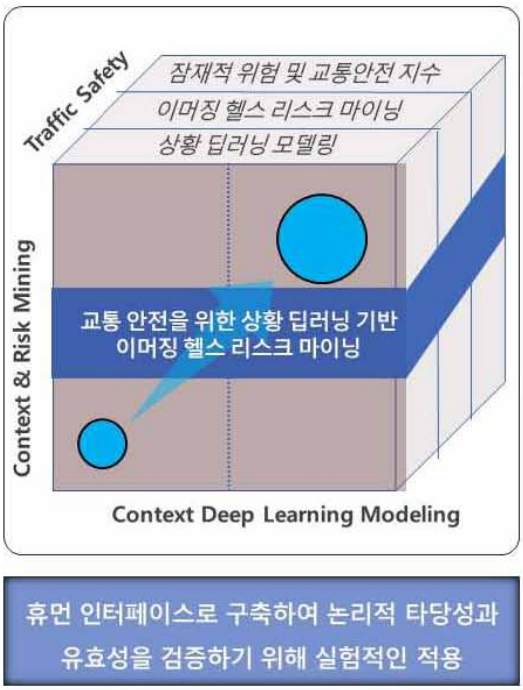
1. 연구개발과제의 개요

1) 연구개발과제의 최종 목표

○ 본 연구의 최종 목표는 과학기술적 측면에서 교통기술과 IT 융복합 기술을 통해 교통안전을 위한 상황 딥러닝 기반 이머징 헬스 리스크 탐지 및 마이닝 예측 기술 개발

- **(주관연구기관)** 교통사고 저감을 위한 상황모션 추적 기반의 행동변화 탐지 및 상황추론, 잠재적 위험에 따른 도로교통 이용자별 교통안전지수, 시공간 결측보정을 통한 LSTM 기반 교통 혼잡도 예측 모델 개발, 빅데이터 기반 교통안전 이머징 리스크 마이닝 예측 기술을 개발

- **(공동연구기관)** 교통사고 데이터 기반 사고발생 빈도 간 상관 분석을 진행하고, 운전자의 내외부 환경 요인 분석을 통한 사고 위험도 예측 모형을 개발하며, 이를 기반으로 한 사고 위험도 예측 시스템 및 모바일 서비스 개발



휴먼 인터페이스로 구축하여 논리적 타당성과 유효성을 검증하기 위해 실험적인 적용

1단계 : 교통안전을 위한 상황 딥러닝 모델 및 교통안전지수 개발

(주관기관)
- 교통안전 및 위험관리 분석
- 위험 예측을 위한 다중회귀 및 상황 딥러닝 모델
- 이머징 리스크 상황 예측
- 상황 모션 추적 기반의 행동변화 탐지 및 상황 추론
- 위험상황 인식 기반 위험패턴 발견 및 교통안전지수 개발

(참여기관)
- 교통사고 데이터 기반 사고발생 빈도 간 상관분석
- 운전자 환경 기반 사고위험도 예측 모형 개발 (내부, 외부)

2단계 : 교통사고 위험도 예측을 위한 멀티미디어 서비스

(주관기관)
- 시공간 결측 보정을 통한 LSTM 기반 교통 잠재적 위험도 분석
- 데이터 특성에 따른 이상치 및 결측 유형 보정
- LSTM 기반 리스트 예측 모델을 통한 임베딩 고도화
- 이머징 리스크 마이닝 인터랙티브 휴먼 인터페이스 개발

(참여기관)
- 도심지 교통사고 위험도 예측 시스템 개발 및 고도화
- 사고 위험도 예측 실시간 모바일 서비스 개발

○ 세부목표

- 교통사고 유발 네트워크 구축을 위한 토픽 모델링과 소셜 빅데이터 분석 기반 이머징 리스크 탐지 및 예측 기술을 개발하여 교통사고 유발 가능성이 있는 잠재적 위험도를 판단함. 위험 상황인식을 위한 지식베이스 분석 알고리즘을 개발하고, 잠재적 위험에 따른 도로교통 이용자별 교통안전지수 서비스 제공

- 국민의 실생활과 밀접하게 관련이 있는 교통안전 이머징 리스크 및 잠재적 위험도를 발견하고 대중적인 관심도가 높은 도로교통 이용자별 교통안전지수를 개발하여 주의사항과 대처방안을 제시하며, 이를 멀티모달 인터랙션 기반의 사용자 인터페이스로 구축하여 논리적 타당성과 유효성을 검증함

• (1차년도) 교통안전을 위한 상황 딥러닝 모델 및 교통안전지수 개발

- 1차년도는 교통안전 서비스를 위해 다양한 요인분석과 상황 딥러닝 예측 모델을 개발하고 이를 기반으로 교통안전지수 개발에 관한 연구 수행

- **주관연구기관** : 교통안전을 위해 다양한 요인분석을 통한 교통안전지수 개발에 관한 연구를 진행하며, 교통안전 및 위험관리 분석, 위험예측을 위한 다중회귀 및 상황 딥러닝 모델, 이머징 리스크 상황 예측, 상황 모션 추적 기반의 행동변화 탐지 및 상황추론, 위험상황 인식 기반 위험패턴 발견 및 교통안전지수 개발에 관한 연구 수행

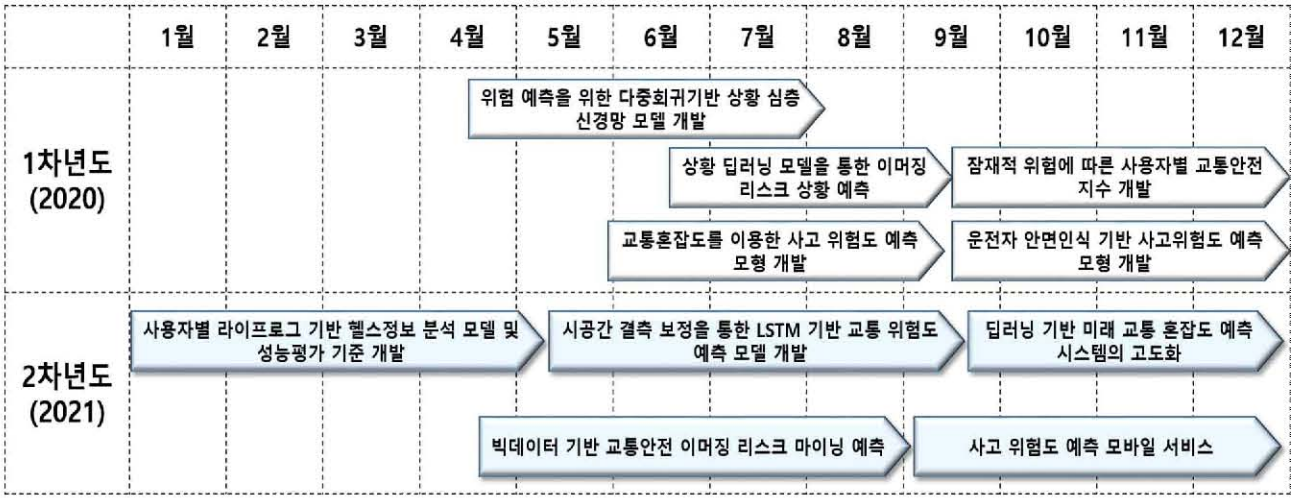
- **공동연구기관** : 교통사고 데이터 기반 사고 발생 빈도 간 상관 분석을 실시하고, 주관연구기관의 요인분석을 반영한 사고 위험도 예측 모형 개발에 관한 연구 수행

• **(2차년도) 교통사고 위험도 예측을 위한 멀티미디어 서비스**

- 2차년도는 1차년도에 개발된 교통사고 위험 예측 모델에 대한 고도화 및 검증, 이를 기반으로 한 인터렉티브 휴먼 인터페이스 소프트웨어 개발 및 사고 위험도 예측을 위한 실시간 멀티미디어 서비스 구현에 관한 연구 수행

- **주관연구기관** : 정확한 위험도 예측 서비스 제공을 위해 시공간 결측 보정을 통한 LSTM 기반 잠재적 교통사고 위험도 분석, 데이터 특성에 따른 이상치 및 결측 유형 보정, LSTM 기반 리스크 예측 모델을 통한 임베딩 고도화 및 이머징 리스크 마이닝, 인터렉티브 휴먼 인터페이스 개발에 관한 연구 수행

- **공동연구기관** : 주관연구기관에서 구축한 사고 위험도 예측 모델들과의 유기적 연계를 통해 사고 위험도 예측 실시간 모바일 서비스 개발 및 도심지 교통사고 위험도 예측 시스템의 체계화 및 고도화에 관련된 연구 수행



2) 연구개발과제의 년차별 목표

○ 1차년도 (2020)

• 위험 예측을 위한 다중회귀기반 상황 심층 신경망 모델 개발

- 차량의 내•외부에서 발생 가능한 잠재적인 위험예측을 위한 상황정보 구성
- 위험 예측을 위한 다중회귀 기반 상황 심층신경망 모델 개발

• 상황 딥러닝 모델을 통한 이머징 리스크 상황 예측

- 위험 상황인식의 정확한 결과를 얻기 위한 공간-시간-주파수 도메인에서 위험 상태감지, 특징추출 및 분류, 결측치 추정 알고리즘 개발
- 평상시 상태 감지와 위험 상황인식의 뚜렷하게 차이나는 공간적 특징을 선택하기 위한 적응적 필터링을 수행하고 시간-주파수 특징맵을 기반으로 시뮬레이션 개발
- 멀티모달 상황 딥러닝을 이용한 3차원 위험 상황인식 모델링

• 잠재적 위험에 따른 사용자별 교통안전지수 개발

- 대화식 정보처리와 서비스를 위한 맞춤형/진화형 의사결정 규칙 개발
- 교통사고 유발 가능성이 있는 상황 모델의 특징 추출, 세분화 모형 개발
- 잠재적 위험 판단을 위한 의사결정 모델을 생성하기 위해 ID3, C5.0, 마이닝 알고리즘 개발
- 마이닝 기반의 분석 및 서비스 모델링/다차원 상황인식 서비스 개발
- 상황패턴, 행동패턴, 상태감지를 이용한 자동식별 의사결정 알고리즘 개발
- 빅데이터에서 마이닝 분석을 통한 도로교통 이용자별 잠재적 위험에 따른 교통안전지수 개발

• 교통혼잡도를 이용한 사고 위험도 예측모형 개발

- 다층퍼셉트론을 이용한 예측모형 구축

• 운전자 안면인식 기반 사고위험도 예측 모형 개발

- 운전자의 실시간 얼굴이미지의 특징을 추출하기 위해 Line-segment feature analysis 알고리즘 개발
- 도로교통 운전자의 안면인식에서 특징점이 되어줄 눈, 코, 입 부분을 도출하는 모듈 개발

○ 2차년도 (2021)

• 사용자별 라이프로그 기반 헬스정보 분석 모델 개발

- 빅데이터 기반 지능형 교통안전 헬스 마이닝 기술
- 사용자별 라이프로그 기반 헬스정보 분석 모델 개발

• 시공간 결측 보정을 통한 LSTM 기반 교통 위험도 예측 모델 개발

- LSTM 기반 교통위험 예측 모델 설계 및 임베딩 고도화
- 데이터 특성에 따른 이상치 및 결측 유형 보정
- 교통사고 심각도 예측을 위한 피드백 기반 앙상블 분류 모델

• 빅데이터 기반 교통안전 이머징 리스크 마이닝 예측

- 빅데이터 기반 이머징 리스크 자동 탐지 및 분석
- 교통사고 유발 네트워크 구축을 위한 토픽 모델링 개발
- 빅데이터 기반 교통안전을 위한 디지털 트윈 모델
- 도로교통 딥러닝 기반 이상 감지 및 마이닝 예측 고도화

• 딥러닝 기반 미래 교통 혼잡도 예측 시스템의 고도화

- 클라우드 자원 이용을 최적화 하기 위한 동적 구성 메커니즘 개발
- 도심지 교통사고 위험도 예측 시스템 개발
- 딥러닝 기반 미래 교통 혼잡도 예측 시스템의 고도화

• 사고 위험도 예측 모바일 서비스

- 실시간 교통정보 제공을 위한 서비스 모듈간 호환 컴포넌트 설계
- 도로교통 가이드라인을 적용한 서비스 인터페이스 개발
- 모바일 서비스 제공을 위한 MVC 패턴 설계 및 사고 위험도 예측 서비스 개발

구분	연구개발목표	핵심 연구개발성과 (연구개발성과물)	현재 국내・외 기술수준	연구개발 목표치	목표치 설정 근거	가중치 (%)
1년차 (2020)	- 위험 예측을 위한 다중회귀기반 상황 심층 신경망 모델 개발	- 논문(SCI) : 2건	- 헬스 빅데이터에서 멀티모달 오토인코더를 이용한 결측치 추론과 심전도 데이터에서 개선된 AnoGAN을 이용한 의사결정 경계 기반 이상치 탐지 모델에 대한 기존의 국내 연구는 전무함 - 딥러닝 기반의 모델은 정밀한 학습을 요구하기 때문에 사용자의 주변환경 및 건강상태, 활동 등에 유연하게 대처하기 어려움	기존대비 10%개선	- 대량의 입력변수를 가지는 데이터 세트는 변수간의 특징을 분석하여 상호 영향력이 높은 변수끼리 구성하는 계획적인 변수 구성 필요 - 일반적인 딥러닝 모델에 데이터마이닝이나 머신러닝 기법을 접목하여 상호간에 발생하는 단점을 보완하고, 장점을 극대화하는 하이브리드 모델링에 대한 연구가 필요	10
	- 상황 딥러닝 모델을 통한	- 논문(SCI) : 1건	- 이머징 리스크 탐지를 위한 감성 유사도	기존대비 10%개선	- 도로교통 이용자의 라이프로그와 상황정보를	10

	이머징 리스크 상황 예측	- 논문(비SCI) : 1건	가중치를 이용한 워드임베딩 기반 교통문서 분류 모델에 대한 기존의 국내연구는 전무함		추론하여 운전 중 갑자기 발생하는 도로교통 이용자에 대한 이머징 헬스 리스크의 문제를 예측 및 발견할 수 있는 방법이 필요함	
	- 잠재적 위험에 따른 사용자별 교통안전지수 개발	- 논문(SCI) : 2건	- 교통안전지수에 관한 국내 연구개발은 한국도로공사에서 차세대 C-ITS 교통사고 예방을 위한 연구만이 제한적으로 이루어짐	기존대비 10%개선	- 상황 추론엔진에서 불필요한 키워드를 제거하고 실시간 스트림 데이터 특성을 고려하여 생성속도 향상 방법이 필요함 - 지식 제공의 효용성과 유용성을 향상시켜 예측하지 못하는 정보를 직관적으로 제공하는 방법이 필요함	10
	- 교통혼잡도를 이용한 사고 위험도 예측모형 개발	- 논문(SCI) : 1건	- 시공간 결측 보정을 통한 LSTM 기반 교통 혼잡도 예측 모델에 대한 기존의 국내연구는 전무함 - 지능형 도로교통 시스템의 일환으로 사용자의 편의 증진을 위해 실시간 교통 서비스에서 도로의 혼잡 가중을 방지하고 교통량을 분산시키기 위한 형태로 이용되기 때문에 신속성에 중점에 두고 있어 정확도가 떨어지는 문제점이 존재함	기존대비 10%개선	- 교통 속도 데이터에는 교통 평균 속도의 흐름을 왜곡시키는 이상치와 수집이 되지 않은 데이터 결측이 존재함. 교통 데이터 특성에 따른 이상치 및 결측을 보정하기 위한 방법 필요	10
	- 운전자 안면인식 기반 사고위험도 예측 모형 개발	- 논문(SCI) : 1건 - 논문(비SCI) : 1건	- 도심 외곽 지역 및 도심지에서 교통 혼잡도와 운전자 안면인식기반 사고위험도에 관한 국내 연구개발은 교통안전을 위한 연구만이 제한적으로 이루어짐 - 운전자 안면인식 모델에 관한 기존의 연구는 있으나, 운전자가 주행시 건강 이상으로 인하여 발생될 수 있는 운전자 건강 위험을 사람의 외부 단서를 통해 정서적 상태를 파악할 수 있는 표정을 통해 건강 상태를 판단에 대한 기존의 국내 연구는 전무함	기존대비 10%개선	- 교통사고 정보 데이터의 편향성 문제를 해결하고 숨겨진 관계를 도출하여 사용자에게 개인 특성과 도로 상황에 맞는 위험 탐지 정보를 제공할 수 있는 방법이 필요함	10
2년차 (2021)	- 사용자별 라이프로그 기반 헬스정보 분석	- 논문(SCI) : 1건	- 라이프로그에서 심전도, 가속도, 호흡 등의 연속된	기존대비 10%개선	- 사용자가 직접 방대한 양의 데이터에서 건강관리 및 질병 예방에 필요한 유의미한	10

	모델 개발		선형구조를 가지는 생체신호 데이터의 건강상태를 판단하는 지표에 대한 연구는 미흡함		정보를 실시간으로 획득하여 이머징 리스크를 예방할 수 있는 방법이 필요함	
	- 시공간 결측 보정을 통한 LSTM 기반 교통 위험도 예측 모델 개발	- 논문(비SCI) : 1건 - 특허 · 출원 : 1건	- 최적의 경로를 제공하여 혼잡한 도로의 혼잡가중을 방지하고 해당 도로의 교통량을 분산시키기 위한 형태로 이용되기 때문에 신속성에 중점을 두어 정확성은 비교적 떨어짐	기존대비 10%개선	- 도로관리를 위한 교통정보 수집 및 가공은 양적인 측면이 아닌 질적인 측면에서 접근이 필요함 - 실시간 교통상황 패턴 예측 및 교통량 예측 연구에서 정확도를 향상시킬 수 있는 방법이 필요함	10
	- 빅데이터 기반 교통안전 이머징 리스크 마이닝 예측	- 논문(SCI) : 3건 - 논문(비SCI) : 1건 - 특허 · 출원 : 1건 - SW 등록 : 2건	- 교통사고 심각도의 판단하는 근거가 부족함 - 블랙박스 시스템에서 수집한 영상 데이터를 분석하여 뚜렷한 특징이 없을 경우 리스크를 분류하는 연구가 미흡함	정확도 10%개선	- 교통사고에 영향을 주는 원인 및 요인들을 파악하여 리스크를 사전에 대비할 수 있는 방법이 필요함 - 교통사고 발생시 효과적으로 처리 및 해결할 수 있도록 객관적인 심각도 평가 지표가 필요함 - 블랙박스 영상에서 도로교통 이머징 리스크를 정확하게 판별하고 분류하기 위한 성능 향상 방법이 필요함	10
	- 딥러닝 기반 미래 교통 혼잡도 예측 시스템의 고도화	- 논문(SCI) : 1건 - 특허 · 출원: 1건	- 교통 데이터의 시계열적 특성에서 특정 값 추출에 대한 어려움이 존재하나 정확도 향상을 위한 전처리 방법에 대해 연구가 미흡함	기존대비 10%개선	- 도로위에서 잠재적인 리스크로 발생하는 이상치 및 결측치에 대한 효과적인 보정 방법이 필요함	10
	- 사고 위험도 예측 모바일 서비스	- SW 등록 : 1건	- 실시간으로 주행 중인 운전자의 상태에 따라 위험도를 예측하기 위한 제약조건이 많기 때문에 정확한 분석에 대한 연구가 미흡함 -	정확도 10%개선	- 실시간으로 주행중인 운전자의 상태를 직관적으로 파악하고, 사전에 대처하여 사고 위험을 방지하기 위해 표정에서 정확한 특징 추출 방법이 필요함 -	10

2. 연구개발과제의 수행 과정 및 수행 내용

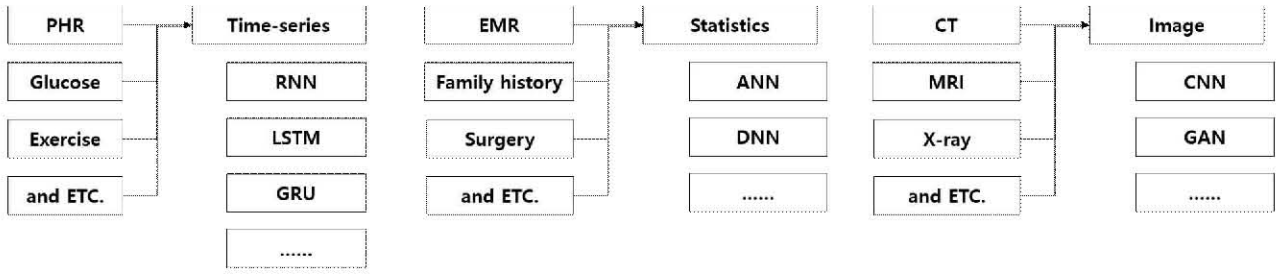
■ 위험예측을 위한 다중회귀기반 상황 심층신경망 모델 개발[1)]

- 의료산업에서 개인건강정보, 전자의무기록, 국민건강정보, 유전체정보 등 활용 가능한 데이터가 증가함에 따라 스마트 헬스가 주목받고 있으며, 모바일 헬스 플랫폼에서 웨어러블 디바이스로부터 다양한 데이터가 장소와 시간에 제약없이 수집되고 있음
- 다양한 스마트 기기를 통해 사용자의 동작, 생체신호, 라이프로그, 날씨, 위치 등이 시계열적으로 축적되고 있으며, 수집되는 데이터는 사용자의 건강과 밀접한 관련이 있으며, 헬스 데이터를 활용하기 위한 다양한 연구가 진행중임

1) J. C. Kim, K. Chung (2020) Hybrid Multi-modal Deep Learning using Collaborative Concat Layer in Health Bigdata, IEEE Access, 8, 192469-192480.

- 웨어러블 디바이스는 산업의 발전에 따라 데이터의 수집 범위가 점점 더 늘어나고 있으며, 이러한 헬스 데이터를 다양한 분야에서 활용하기 위해서 머신러닝, 딥러닝, 데이터마이닝 등의 방법을 이용한 헬스 모델들이 개발되고 있음
- 국민건강영양조사 데이터는 건강설문조사, 검진조사, 영양조사를 통해 수집된 데이터이며, 건강설문조사를 통해 개인의 질병기록, 생활습관, 진료기록 등을 수집하고, 정기검진을 통해 혈압, 맥박, 혈당, 신장, 체중 등을 수집함. 영양조사는 개인의 물 섭취량, 식사량, 식사 빈도, 영양제 복용여부 등 설문을 통해 수집함. 이는 일반적인 개인건강기록과 거리가 있지만, 헬스와 관련된 다양한 변수를 포함하고 있어 잠재적 가치가 높으며, 약 600개의 변수로 구성되어 있고, 데이터마이닝 기법을 통한 분석으로 각 변수들의 규칙이나 관계를 추론할 수 있음
- 4차 산업혁명에 따라 수많은 데이터들이 수집되고 있으며, 이를 이용하는 스마트 헬스 산업이 지속적으로 발전하고 있음. 스마트 헬스는 다양한 기기와 정보통신공학의 보급으로 많은 양의 데이터를 활용할 수 있게 되었으며, 광범위한 데이터들을 활용하기 위해서 다양한 딥러닝 모델이 개발되고 있으며, 다양한 경로에서 수집되는 데이터들의 통합이 요구되고 있으며, 일상생활에서 활용되는 모델은 한 개의 변수에 대해서 time-step이 중복되어 수집될 가능성이 존재함
- 각 변수에 따라 관측되는 시점이 다를 경우 하나의 입력으로 처리하기 어려움. 데이터마이닝과 딥러닝을 이용한 모델은 사용되는 학습 데이터, 변수 구성, 목표 등에 따라 다양하게 생성됨. 스마트 헬스 데이터는 이미지나 시계열, 비시계열 형태로 이루어져 있으며, 시계열 데이터는 수집되는 주기가 일정하고 빠른 데이터를 의미함
- 비시계열 데이터는 주기가 일정하지 않은 데이터를 의미하고, 이미지 데이터는 MRI, CT, 초음파, X-ray 등을 나타냄. 사람의 건강에 다양한 변수들이 영향을 주고 있으며, 각 변수들의 특성이 다르기 때문에 통합 및 연결 문제가 존재하고, 딥러닝 모델을 학습하기 위해서는 많은 양의 학습 데이터가 요구됨
- 헬스 데이터의 발생 주기가 현저하게 상이하고, 어느 한 시점을 기준으로 학습 데이터를 병합하여 구성하게될 경우 빠른 주기를 갖는 데이터는 일정 부분의 정보손실이 발생하고, 느린 주기를 갖는 데이터의 경우 트랜잭션에 동일한 값으로 채워지기 때문에 모델을 통해 출력된 결과의 영향력이 낮다는 문제점이 있음
- 시간적인 연속성을 가진 데이터를 시계열 데이터라고 하며, GRU나 LSTM과 같은 RNN을 이용하여 모델링하고, 사용자에 따라 대량의 데이터 수집이 가능하며, 사용자에 따른 학습 데이터 구성이 가능함. IoT와 개인건강기기가 보급됨으로써 개인의 건강과 관련된 데이터들의 수집이 용이해졌으며, 헬스 데이터 생산의 주체가 기관에서 개인으로 이동하는 원동력이 되었으며, 현재 다양한 개인용 헬스 장비에서 사용자의 건강 관련 데이터를 수집하여 분석함으로써 이를 활용한 스마트 헬스 서비스들이 개발되고 있음
- 시계열 데이터에 대한 딥러닝 모델은 사용자의 기록으로부터 현재 혹은 미래의 값을 예측하는 방법으로 사용하고 활동량, 심박수, 혈압 등의 변화를 예측하여 사용자 건강을 모니터링하고 부정적인 영향을 주는 잠재적인 변수들을 도출함
- 비시계열 데이터는 일반적인 데이터마이닝 기법을 이용하여 모델링하고, 각 사용자에 대해 하나의 트랜잭션이 존재하며, 다수의 사용자의 데이터를 수집해야 함. 건강검진, 국민건강영양조사와 같은 EMR에서 나타나는 변수들이 비시계열 데이터로 활용됨. 대량의 데이터를 각종 데이터 마이닝 기법으로 분석하여 나타나는 결과를 보편적인 서비스에 활용함. 주로 의사결정 나무, 연관규칙, 심층신경망 등이 사용되기 때문에 실제 사용자의 건강 상태를 예측하거나 추론함

• 건강 데이터에 따른 딥러닝 방법



- 이미지 데이터는 비시계열 데이터와 동일하게 여러 사용자의 데이터를 활용하며, X-ray, MRI, CT 등의 의료 이미지를 학습시킴으로써 사용자들의 건강 상태를 예측함. 이미지 데이터에서 많이 사용되는 CNN 모델

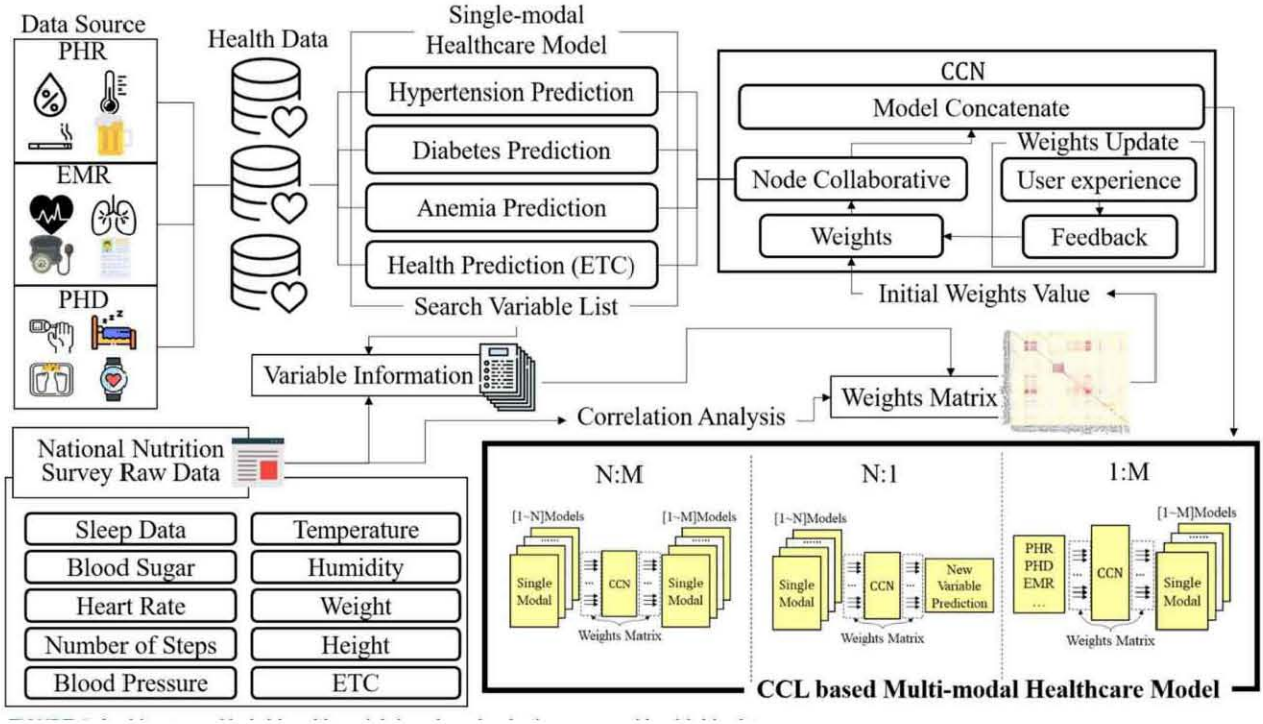
을 기반으로 다양한 신경망 모델들이 사용됨

- 스마트 헬스에서 데이터를 병합하여 학습 및 분석에 활용될 때, 변수의 부재나 특징 추출 등의 문제가 발생하기 때문에 각기 다른 특징을 갖는 헬스 데이터에 따라 각각의 싱글-모달로 구성함. 각 모달에 포함되는 헬스 변수에 따라 모델링의 결과나 학습효율의 편차가 존재함
- 데이터 특징에 따른 싱글-모달 모델들을 하나의 멀티-모달로 구성함으로써 각 싱글-모달들의 단점을 보완할 수 있고, 사용자의 상황을 고려하여 다른 싱글-모달 모델이 요구되며, 각각의 싱글-모달들을 통합한 일반적인 멀티-모달 모델은 사용자들의 각기 다른 상황을 고려하지 못하는 문제점이 존재하며, 이에 따라 싱글-모달 모델들을 효율적으로 연결하는 방법이 필요함

■ 헬스 빅데이터에서 Collaborative Concat Layer를 이용한 하이브리드 멀티-모달 딥러닝[2)]

- 헬스 데이터는 디바이스 혹은 관측대상에 따라 여러 경로를 통해 수집되어 발생 주기, 단위, 형태가 상이하기 때문에 하나의 데이터 집합으로 통합이 어려운 문제가 있고, 데이터가 여러 기관과 개인 디바이스에 분산되어 있음
- 헬스 데이터의 다양한 변수나 개인정보 보호 문제로 인해 통합이 어렵고, 수집 경로나 특징에 따라 시계열, 비시계열, 이미지 등 다양한 형태의 데이터들이 수집됨
- 건강은 여러가지 요인에 의해 영향을 받기 때문에 다양한 종류의 데이터를 이용하여 예측하며, 예측하고자 하는 목표변수에 따라 각기 다른 모델들이 필요하기 때문에 헬스 모델은 다양한 형태와 변수로 구성됨
- 헬스 데이터는 한 곳에서 수집되는게 아니기 때문에 여러 곳을 통해 수집되는 멀티-모달리티의 특징이 존재하고, 각 데이터의 특성에 따라 싱글-모달 모델로 주요한 특징들을 추출하여 결합하는 멀티-모달 딥러닝 모델이 주목받고 있음. 딥러닝을 이용한 헬스 모델은 기존의 머신러닝이나 데이터마이닝에 비해 높은 퍼포먼스를 나타내기 위해 많은 양의 데이터를 이용한 신경망 학습이 요구됨
- 딥러닝 모델은 다계층의 레이어로 구성된 신경망으로, 딥러닝의 초기 가중치를 통해서 레이어에 구성된 노드들 간의 연결이 이루어지고, 각 노드에서 가중치를 업데이트하여 다음 노드로 결과값을 전송함. 사용자가 입력한 값에 대해서 정답 값에 근접하는 값을 출력하도록 모델의 업데이트를 반복하여 입력된 데이터의 특징을 학습함

• 하이브리드 멀티-모달 딥러닝 헬스 예측 모델의 구조



2) J. C. Kim, K. Chung (2020) Hybrid Multi-modal Deep Learning using Collaborative Concat Layer in Health Bigdata, IEEE Access, 8, 192469-192480.

• 히든노드 x_{10}의 계산 방법

$$x_{10} = x_{00} \times w_{00} + x_{01} \times w_{01} + b$$

• 출력노드 Y_{00}의 계산 방법

$$y_{00} = x_{10} \times w_{10} + x_{11} \times w_{11} + b$$

- b는 바이어스를 나타내고, 2×2×2형태의 간단한 신경망으로 2개의 입력과 2개의 출력으로 구성함
- 딥러닝은 계속되는 학습을 통해 가중치 업데이트를 반복하여 최적의 값을 탐색하고, 모델에서 선정되는 초기 가중치는 학습에 영향을 미침. 초기 가중치는 일반적으로 랜덤한 값을 사용하거나 오토인코더를 사용해 사전학습을 진행함

- 일반적으로 신경망 구조에서 입력값에 결측치가 존재하는 경우, 모델의 정확도가 낮아지는 문제점이 존재함. 단기적으로 입력값이 결측될 경우 평균값이나 중앙값 등을 통해 값을 대치할 수 있지만, 이러한 방법은 스마트 헬스에서 입력값이 장기적으로 결측치로 존재할 때 신경망 모델을 재사용하기 어렵다는 문제점이 존재함
- 나이, 키, 혈당, 체중 등의 개인적인 상황 정보들을 통해 만성질환을 예측하는 신경망 모델에서 혈당에 해당하는 데이터들은 개인건강기기를 통해 수집이 이루어지고, 혈당측정기를 소유하고 있지 않을 경우 혈당 데이터의 장기적인 부재가 발생할 수 있음
- 상황 마이닝을 통해 도출된 각 변수 간의 관계성을 통해 혈당을 예측함으로써 혈당 데이터가 존재하지 않는 데이터를 이용하여 신경망 모델을 사용할 수 있고, 각각의 신경망 모델을 연결하여 멀티-모달 신경망 모델로 확장이 가능함

• 신경망 구조에서 사용되는 변수정보

Variable Name	Code	Average	Notice
age	age	47.26	User's age
subjective health	D_1_1	2.81	User's own health status
hypertension diagnosis	DI1_dg	6.50	Hypertension diagnosis result by doctor
dyslipidemia	DI2_dg	6.92	Dyslipidemia diagnosis result by doctor
stroke	DI3_dg	7.85	Stroke diagnosis result by doctor
myocardial infarction	DI5_dg	7.93	Myocardial infarction diagnosis result by doctor
weight	HE_wt	62.32	User recent weight
kidney	HE_ht	162.28	User recent height
waist circumference	HE_wc	80.63	User recent waist circumference
body mass index	HE_BMI	23.55	User's latest body mass index
.....	.....	.....	.....

- 딥러닝과 데이터마이닝을 함께 사용함으로써 일반적인 모델에서 발생하는 문제점을 해결한 하이브리드 모델을 생성할 수 있으며, Collaborative Concat Layer (CCL)는 연관 마이닝을 기반으로 연결하는 방법으로 가중치 행렬을 이용하여 각 변수간의 관계성을 통해서 결측된 값을 예측함

▪ 건강 변수에 의한 벡터화를 사용한 가중치 행렬

- 가중치 행렬에서 Collaborative Concat Node의 초기 가중치를 저장하기 위한 행렬을 사용하여 국민건강영양

조사 원시자료를 기반으로 진행하고, PHQ에 대한 정보를 추가로 수집하여 데이터 집합을 구성함

- 국민건강영양조사 데이터는 다양한 질환에 대한 통계자료로써 사람들의 상황에 따라 결측값이 다량 포함되어 있음. 결측값은 출력되는 모델의 결과에 부정적인 영향을 미치기 때문에 전처리를 진행함
- 국민건강영양조사 원시 데이터에서 헬스 서비스와 직접적인 연관이 있는 196개 변수를 사용함
- Collaborative Concat Node를 산출하기 위해 변수의 평균값을 국민건강영양조사 원시자료 이용지침서를 이용하여 데이터 집합을 구성함
- 전체 트랜잭션에서 데이터 정제와 가공을 통해 추출한 10,488건의 데이터를 이용하고, 무작위로 70%인 7,342건의 데이터를 분석에 사용하며, 나머지 3,146건의 데이터를 실험 및 평가에 사용함

• 국민건강영양조사 원시자료

No.	age	D_1_1	DI1_dg	DI2_dg	.....	HE_wt	HE_ht	HE_wc	.....
1	76	2	1	0	.....	59.7	166.7	81.4	.....
2	39	2	0	0	.....	62.7	171.7	75.7	.....
3	35	2	0	0	.....	58.3	169.5	77.6	.....
4	71	3	0	0	.....	53.1	169	68.1	.....
5	68	3	1	1	.....	61.3	158.1	89.8	.....
6	28	2	0	0	.....	53.9	156.7	75.9	.....
7	45	3	0	0	.....	48.8	160.7	66.9	.....
8	70	3	1	0	.....	53.7	145.3	86	.....
9	37	4	0	0	.....	55	159.8	71.7	.....
10	67	2	0	1	.....	73.9	165.7	86.9	.....
11	57	3	0	0	.....	61.8	158	89.1	.....
12	46	3	0	0	.....	68.3	175.8	86.3	.....
13	33	2	0	0	.....	59.3	168.5	73	.....
14	41	3	1	0	.....	80.1	173.1	89.5	.....
15	23	2	0	0	.....	64.5	162	74.7	.....
16	73	4	1	1	.....	52.8	151.3	83	.....
17	78	2	1	1	.....	56.3	155	77.6	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....	.....

- 전처리된 데이터를 변수간의 상관관계 분석을 통해 가중치 행렬을 구성함. 상관관계 분석은 변수 간의 상관정도를 수치적으로 표현하는 데이터 마이닝 기법 중 하나임

• 구성한 변수에 대한 가중치 매트릭스

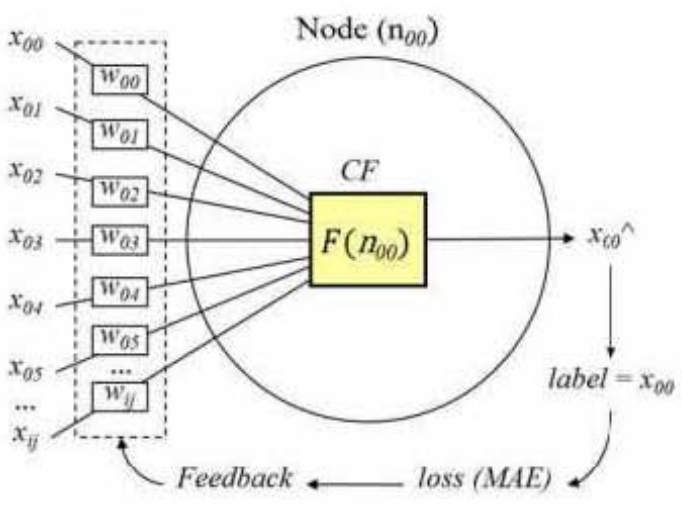
key1 / key2	age	D_1_1	DI1_pr	DI2_pr	DI3_pr	...	N_DIET	N_DIET_WHY	N_WATC
age	0.00	0.30	-0.48	-0.32	-0.15	...	-0.06	-0.08	-0.06
D_1_1	0.30	0.00	-0.23	-0.20	-0.13	...	-0.05	-0.07	-0.04
DI1_pr	-0.48	-0.23	0.00	0.35	0.17	...	0.06	0.08	0.02
DI2_pr	-0.32	-0.20	0.35	0.00	0.10	...	0.10	0.12	-0.00
DI3_pr	-0.15	-0.13	0.17	0.10	0.00	...	0.02	0.03	0.01
DI5_pr	-0.10	-0.08	0.09	0.07	0.08	...	0.02	0.03	-0.01
DI6_pr	-0.15	-0.11	0.13	0.12	0.06	...	0.02	0.03	-0.00
...	...	...	...	...	...	...	...	...	...
N_DIET	-0.06	-0.05	0.06	0.10	0.02	...	0.00	0.91	-0.06
N_DIET_WHY	-0.08	-0.07	0.08	0.12	0.03	...	0.91	0.00	-0.07
N_WATC	-0.06	-0.04	0.02	-0.00	0.01	...	-0.06	-0.07	0.00

- 행렬에서 가중치를 검색하기 위해 가로축은 key1, 세로축은 key2로 사용함. 데이터베이스가 갱신될 때 가중치 값이 변화할 수 있고, 상관관계 분석결과 2개의 변수 상관정도가 동일할 경우, 1로 나타남. 그러나 프로그램 구현에서 연산에서 편의를 위해 key1과 key2가 동일할 경우 0으로 저장하여 처리함

■ **Collaborative Node 구조**

- Collaborative Node는 상관계수를 가중치 초기값으로 사용함. 상관계수는 신경망 모델을 통해 학습된 가중치는 아니지만, 데이터마이닝과 머신러닝에서 일반적으로 사용되는 가중치로써 이를 사용한 다양한 예측이 가능함. 신경망에서 노드의 연산은 입력값과 가중치를 활성화 함수에 적용하여 결과값을 출력함

• Collaborative Node에서 Node(n_{00})



- Node(n00)은 CN에 연결된 딥러닝 모델의 00번째부터 ij번째까지 출력값을 이용하여 x_{00}^-를 출력함
- Node(n00)에서 w01은 입력변수 x00과 x01의 상관계수를 나타내고, 상황정보 행렬에서 탐색한 key값을 이용하여 가중치 행렬에서 가져오고, key1값은 CN 노드의 출력에 따라 정해지고, key2값은 연결된 딥러닝 모델의 출력에 의해 정해짐
- Node(n00)에서 key1은 x_{00}^-이고, key2는 x_{ij}을 나타냄. 가중치 w00은 입력변수 x00과 x00의 상관계수를 나타내고, key1과 key2가 동일할 경우 같은 변수로 가중치는 0으로 설정함
- Node(n00)의 출력이 x_{00}^-일 때 x00와 x_{00}^-는 동일한 의미를 가진 변수이기 때문에 x_{00}^-의 예측에서 x00를 제외하기 위함. Node(n01)일 경우 w00은 x01과 x00의 상관계수를 나타내고, Node(n01)에서 w01은 입력변수 x01과 x01의 상관계수로 0이 됨

• Node(n_{00})의 활성화 함수

$$f(n_{00}) = \hat{x_{00}} = \overline{x_{00}} + \frac{\sum_{i,j=0} (x_{i,j} - \overline{x_{i,j}}) \times w_{i,j}}{(\sum_{i,j=0} |w_{i,j}|)}$$

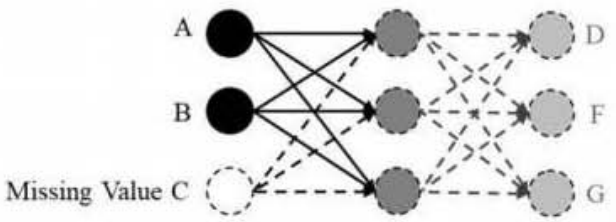
- x_{00}^-는 노드 Node(n00)의 출력을 나타내고, w는 가중치, b는 바이어스를 나타냄

■ **Collaborative Concat Layer를 이용한 하이브리드 멀티-모달 딥러닝**

- 스마트 헬스는 각 사용자의 상황을 고려한 개인화되고 적합한 서비스 제공을 요구하고 있음. 사용자가 소유하고 있는 디바이스나 환경에 따라서 활용할 수 있는 데이터가 상이하고, 사용자의 상황에 변화할 경우 헬스 모델의 정상적인 동작이 어려움

- 입력변수 C가 결측치일 경우 하위에 있는 모든 노드에 부정적인 영향을 주게 됨. 이와 같이 모델의 성능이 낮아지는 문제점을 보완하기 위해 각각의 헬스 모델이 상호 보완 가능하도록 구성해야 함

• 사용자 상황 변화에 따른 헬스 모델의 오류



- Collaborative Node를 사용하여 부재중인 변수를 대치할 수 있음. 각 노드 간의 연산을 통해 결측된 변수의 값을 계산하고, 변수가 결측되지 않은 경우 그대로 작동해야 함
- 여러 개의 CN을 사용하여 여러 가지 오류로 인해 발생하는 헬스 데이터 결측치에 대처하는 Layer 구성함

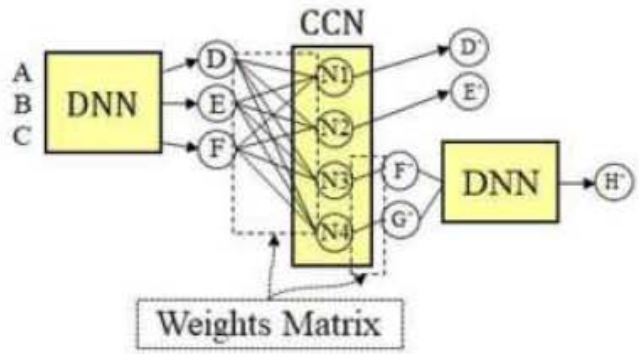
• Collaborative Concat Layer 생성 알고리즘

Algorithm 1 Collaborative Concat Layer Generation

Previous Model n, Post Model m
$x'_{n,i}$: Model n's Output i
$y'_{m,i}$: Model m's Input j
k: number of variables
CNi: Collaborative Concat Layer's Node i

1 Search Output Variables in Model()
2 Search Input Variables in Model()
3 Create CNi
4 CNi input = for(i, 1:Kn) for(j, 1:Km)
5 initial weight $w_k(x'_{n,i}, y'_{m,i})$ = similarity$(x_{n,i}, y_{m,i})$
// from Weights Matrix
6 parameter $P_{n,i}$ = average $x_{n,i}$// from variable information
7 CNi output = for(i, 1:K_n) for(j, 1:K_m)
8 if($x'_{n,i} == y'_{m,i}$)
9 if($x'_{n,i}$! = *Null*)$y'_{m,i} = x'_{n,i}$
10 else $y'_{m,i}$ = *Prediction*

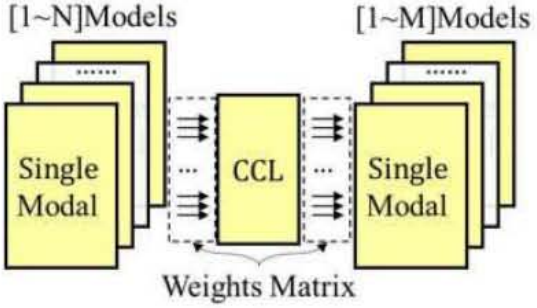
• Collaborative Concat Layer 생성 프로세스[3)]



- 입력값 A, B, C를 통해 출력 D, E, F를 예측하는 DNN 보델과 입력값 F, G를 통해 H를 예측하는 DNN 보델을 연결하였고, Collaborative Concat Layer에서 초기 생성 시 연결되는 DNN의 출력변수와 입력변수를 변수 행렬에서 검색하고 가중치 행렬을 적용하였음
- G′ 는 전위위 DNN의 출력에서 없는 변수로 N4에서 연결되는 후위 DNN의 입력변수를 key1로 하여 가중치 행렬에서 검색하여, 변수 F나 G의 값이 없을 경우 CN을 통해 대치된 값이 후위 DNN 보델의 입력값으로 들어가는 구조임

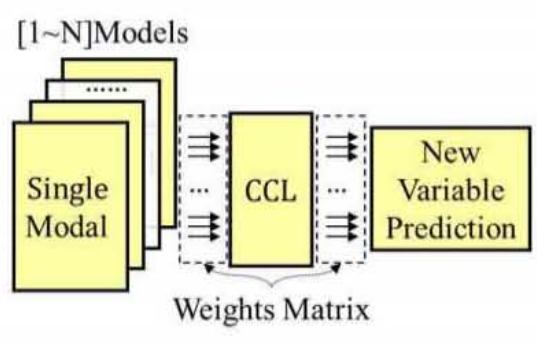
3) 김주참, 예측 정확도 향상을 위한 협력적 레이어 기반 하이브리드 멀티-모달 딥러님, 경기대학교 컴퓨터과학과, 박사학위논문, 2020.12.

• N-CCL-M Multi-modal Model



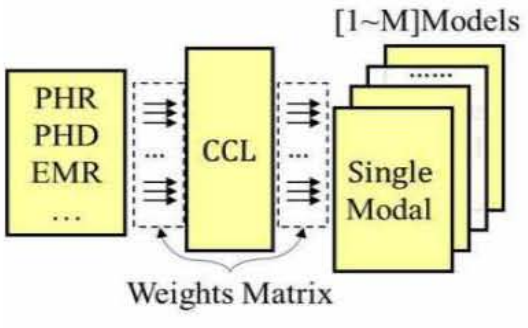
- 전위 싱글-모달 모델 N개와 후위 싱글-모달 모델 M개가 CN을 이용하여 연결되는 구조이며, 전위에 있는 모델의 출력값과 가중치 행렬을 이용하여, 후위에 있는 모델의 결측치로 존재하는 입력변수를 대치하여 연결함

• N-CCL Multi-modal Model



- N-CCL 모델에서 CCL은 전위의 있는 싱글-모달 모델의 출력값과 가중치 행렬을 이용하여 새로운 변수에 대한 예측이 가능하며, 가중치 매트릭스에 존재하는 변수에 한에서 가능하지만, 행렬에서 새로운 데이터가 입력될 경우 더 많은 변수에 대해 예측이 가능함

• CCL-M Multi-modal Model



- 후위에 싱글-모달 모델이 위치하는 구성으로, 사용자의 특정 변수가 결측값으로 존재할 경우, CCL에서 예측한 값으로 대치하여 후위에 있는 싱글-모달 모델의 활용이 가능함
- N-CCL-M, N-CCL, CCL-M Multi-modal Model을 통해 각기 다른 사용자의 상황에 적합하게 변경하여 활용할 수 있으므로 사용자의 상황에 따른 유연한 모델의 구성을 통한 유연한 이머징 헬스 서비스가 가능함

■ 헬스 빅데이터에서 멀티-모달 Denoising 오토인코더를 이용한 결측치 추론[4)]

- 헬스 빅데이터는 인간의 건강과 관련된 모든 데이터를 의미하며, 정보통신기술의 발전으로 헬스케어 분야에서 데이터 수집이 용이해지면서 정부기관이나 병원뿐만 아니라 개인에 의해서도 헬스 데이터가 수집됨
- EMR, PHR, 라이프로그 등의 데이터는 공통되는 변수들이 존재하며, 사용자에 따라 다양한 변수들로 구성되어 있음
- 개인유전정보는 환자 개인에게 적합한 최적의 의료서비스를 제공하는 정밀의료의 핵심 정보로써 치료비용을 감소하고 효율을 높이는데 주된 역할을 하고 있음
- 유전공학의 발전으로 다양한 민간 의료업체가 등장하면서 유전자 검사비용이 저렴해지고, 병원을 통하지 않고 분석 의뢰가 가능해짐. 개인유전정보는 정밀의료와 같이 개인에 집중한 의료서비스에서 활용됨

- 개인건강기록(PHR)은 다양한 스마트 헬스 디바이스를 통해 수집되는 데이터와 개인이 스스로 기록하는 모든 의료행위를 포함하는 데이터이며, 개인이 건강 데이터를 직접 수집하여 열람하고 관리하는 데이터이고, 개인의 디바이스나 상황에 따라서 다양하게 구성하고 통합하여 활용함
- 전자의무기록(EMR)은 의사가 진료 기록을 업무 자동화 시스템에 입력하고, 관리되는 여러 진료 정보들을 나타내며, 의료 빅데이터 처리가 가능해졌고, 데이터마이닝 기술과 접목하여 정밀의료의 근간이 되고 있음

- 국민건강정보는 국민의 건강을 증진하기 위해서 국가가 수집하는 데이터를 나타내며, 국민건강영양조사는 일반적인 개인건강기록 데이터는 아니지만, 개인건강기록에 포함되는 다량의 데이터를 포함하고 있기 때문에 잠재적인 가치가 매우 높음
- 이는 약 600여개의 변수로 구성되어 있음. 국민건강영양조사 데이터를 사용해 도출한 모델은 개인건강기록에서 활용성이 높음
- 국민건강영양조사는 다양한 경로를 통해서 수집된 데이터로써 멀티-모달리티적인 특성을 지니고 있음. 데이터가 수집되는 경로에 따라 각각의 모달을 통해 수집. 통합하는 과정에서 사람들의 상황이 상이하여 결측치가 다량 포함
- 결측치는 여러 모델의 학습 결과에 부정적인 영향을 미치고, 모델의 학습이나 분석에서 오류가 발생할 수도 있음. 또한 다양하고 불규칙하게 발생되기 때문에, 결측치에 해당하는 트랜잭션 값을 모두 삭제할 경우 많은 양의 정보가 손실됨
- 결측치를 추론하는 방법은 평균값, 최빈값, k-NN, Support Vector Machine(SVM), 딥러닝 등이 존재함

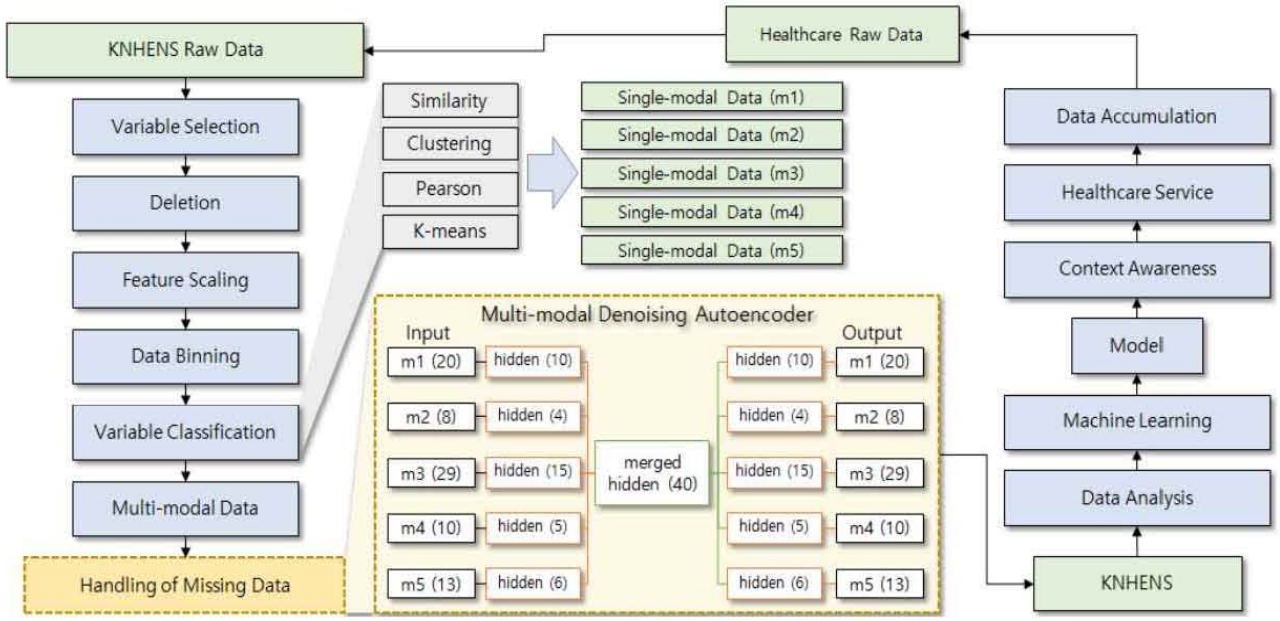
- 평균값이나 중앙값을 이용한 추론은 변수에서 계산을 통해 결측치 추론. 각각의 열은 독립적으로 결측치를 추론하며, 수치형 데이터에서만 사용이 가능하며, 계산이 간단하고 빠르게 결측치를 추론할 수 있음. 중앙값이나 평균값을 이용한 방법은 여러 변수 간의 관계를 고려하지 않기 때문에 범주형 데이터에서 정확도가 낮아지는 문제가 있음
- 최빈값을 이용한 결측치 추론은 매우 간단하게 작동하며, 수치형 데이터보다는 범주형 데이터에서 주로 사용됨. 최빈값 또한 여러 변수간의 관계를 고려하지 않기 때문에 가장 빈번하게 관측되는 값으로 추론할 경우 데이터가 편향될 수 있음. 데이터가 편향될 경우 학습 및 분석 결과가 의도와 다르게 나타나며, 한 변수의 집중된 상태로 나타날 수 있음
- k-nn을 이용한 추론 방법은 결측치가 발생한 관측값과 가장 가까운 k개의 이웃을 찾고, 그 이웃들의 가중평균값으로 결측치를 추론하는 방법이며, 간단한 분류에 사용되는 알고리즘으로 Feature similarity를 이용하여 새로운 데이터의 값을 예측함
- k-nn은 일반적으로 최빈값, 평균값 등의 방법 보다 더 높은 정확도를 가지지만, 전체 학습 데이터를 메모리에 저장해야 작동하며, 계산량이 크다는 단점이 존재함. 지정된 k값에 따라서 결과가 변동되기 때문에 적절한 k값을 찾아야 함
- SVD를 이용한 추론은 선형대수학에서 행렬을 대각화하여 입력 x와 유사한 출력값 $\hat{x}$를 통해 결측치를 예측하는 방법이며, 결측값을 0이나 column의 평균값으로 초기화한 다음 임계값으로 수렴할 때 까지 k개의 가

4) J. C. Kim, K. Chung (2020) Multi-Modal Stacked Denoising Autoencoder for Handling Missing Data in Healthcare Big Data, IEEE Access, 8, 104933 - 104943.

장 중요한 고유 변수의 선형 조합을 반복하여 추정

- SVD는 모든 행렬에 대해 적용이 가능하고, 고유값 분해를 해서 생성된 직교행렬을 이용하여 입력값 x와 유사한 $\hat{x}$를 출력함. x에서 결측된 부분을 $\hat{x}$의 값으로 추론하여 결측치를 처리함

● 멀티-모달 Denoising 오토인코더를 이용한 결측치 추론 방법의 구성도



- 딥러닝을 이용한 추론은 신경망을 학습하여 생성된 가중치를 통해서 결측치를 예측하는 방법이며, 신경망 구조에 따라 표현이 가능하며, 결측치 추론을 위한 대표적인 구조로 오토인코더가 있음

- 신경망에서 입력노드보다 히든노드가 작을 경우 입력 데이터를 압축하고, 특징을 추출 가능하고, 일반적으로 사전학습에서 활용되지만, 입력값을 복구하려는 특징을 이용하여 결측치에 대한 복구가 가능함
- 트랜잭션에 결측치가 많이 포함되어 있을 경우 데이터를 전처리하는 방법에 따라서 분석 결과가 현저히 상이하게 나타남. 데이터의 결과에 결측치가 미치는 영향을 줄이기 위해 적절한 처리방법이 필요하기 때문에 멀티-모달 구조의 데이터에서 발생하는 결측치를 학습된 오토인코더를 이용하여 결측치를 추론함

■ **질병관리청 국민건강영양조사 데이터 전처리**[5]

- 원시 데이터로 질병관리청 국민건강영양조사 중 검진조사, 건강설문조사, 영양조사 데이터를 활용. 각 연도별로 조사되는 범위가 달라지기 때문에 공통된 변수만을 활용하고, 컬럼 스캔 결과 결측치가 20%가 넘는 변수는 제외함
- 결측값은 아니지만 데이터 분석 결과에 영향을 미칠 수 있는 부류가 존재. 비해당은 많은 변수에 포함되지만 14,688건 중 이를 대량(3,000건 이상)으로 포함하고 있는 변수들은 제외함
- 실험데이터 14,688건에서 범주형 변수에 대해 1을 증가하여 0(없음)을 1(없음)으로, 1(있음)을 2(있음)으로 값을 변화. NULL, NA, 무응답(9, 99, 999) 등의 값들은 오토인코더에서 결측치를 나타내는 0으로 전처리 진행함

- 연속형 변수의 범위가 상이할 경우 가중치가 0으로 수렴하거나 값이 지나치게 커지는 문제 발생. 변수간의 scale이 상이할 경우 가중치가 불안정하게 학습될 수 있어 Feature scaling이 필요하기 때문에 데이터 Binning을 통해 변수들을 범주형으로 변환하며, 동시에 Feature scaling을 진행. 변수의 최솟값을 0%로 최댓값을 100%로 하여 9개의 구간으로 범주화
- 실험데이터에서 결측치를 0으로 정했기 때문에 1부터 9까지로 binning 진행함

● 질병관리청 국민건강영양조사 데이터에서 연속형 변수의 binning과 Feature scaling 결과

5) 질병관리청 국민건강영양조사, https://knhanes.kdca.go.kr/knhanes/.

Binning	0	1	2	3	4
Normalization	N/A	0–0.1	0.1–0.2	0.2–0.3	0.3–0.4
Binning	5	6	7	8	9
Normalization	0.4–0.5	0.6–0.7	0.7–0.8	0.8–0.9	0.9–1.0

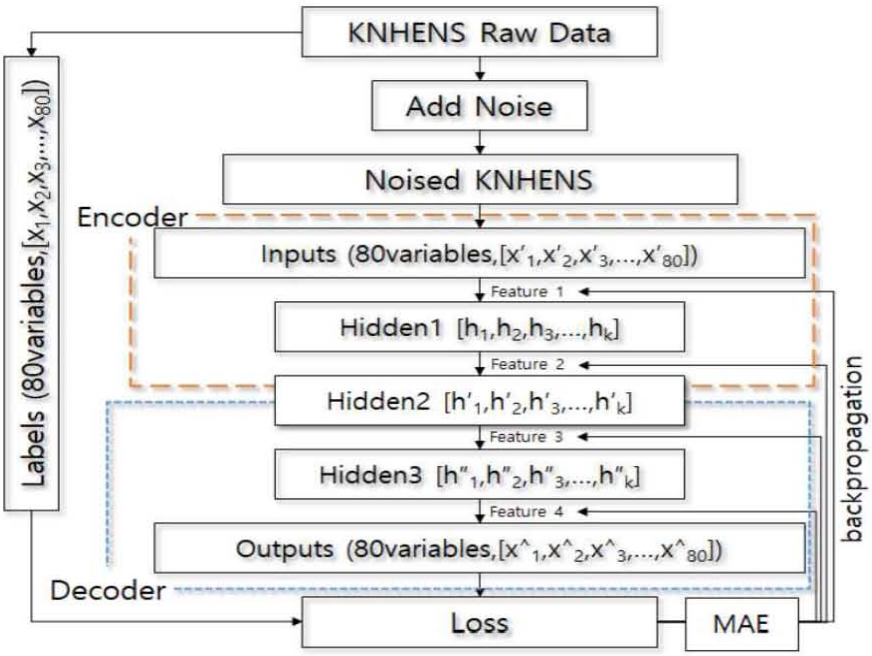
■ **Denoising 오토인코더를 이용한 결측치 추론**

- Denoising 오토인코더는 일반적인 오토인코더의 학습 방법을 변형시킨 모델이고, 잡음이 없는 입력 데이터에 임의로 잡음을 추가한 데이터를 입력값으로 함. 이는 잡음이 없는 원본 데이터를 복원 목표로 학습하는 오토인코더를 나타내고, 입력값에 무작위로 노이즈를 포함하여 원본으로 복원하는 과정을 학습함
- Denoising 오토인코더는 데이터 입력 전에 원본 데이터에서 원래 값을 무작위로 선택하여 0으로 바꾸고, Denoising 오토인코더 학습에서 노이즈는 0으로 추론하여 결측치가 발생했을 경우 이를 원본 데이터로 복원하여 결측치가 발생할 경우 0으로 입력하고, 학습된 신경망을 거쳐 0이 아닌 예측값으로 추론함

■ **Denoising 오토인코더 기반 결측치 추론**

- 원본 데이터는 노이즈 추가 단계를 거치고, 이를 트레이닝 데이터로 변환하고, 원본 데이터를 라벨로 하여 신경망의 출력값의 오차를 이용하여 가중치 학습

• Denoising 오토인코더 기반 결측치 추론 프로세스



- 히든 레이어가 5개(64, 32, 16, 32, 64)로 구성된 오토인코더는 80(입력) → 64(히든) → 32(히든) → 16(히든) → 32(히든) → 64(히든) → 80(출력)으로 80부터 16까지가 인코더이고 16부터 80까지가 디코더로 사용함. 원시 데이터에서 트랜잭션은 약 25%의 결측치를 가지고 있으며, 학습에서 노이즈 factor는 0.05부터 0.30까지 0.05씩 증가하며 진행하고 활성화 함수는 Relu(Rectifier Linear Unit), optimizer는 adam을 사용함

- 파라미터가 많을수록 더 많은 특징 정보를 포함하고 있으며, 학습결과 64개의 노드를 가진 히든레이어 하나로 구성할 경우 정확도와 loss가 가장 우수하게 나타남
- 파라미터의 수가 많을수록 실제 시스템의 연산성능이 저하되지만 두 모델의 파라미터 차이는 480개로 큰 차이가 없어 정확도가 더 우수한 80-64-80 모델을 선택하고, 모델의 정확도는 0.9421로 대부분의 결측값이 의미있는 값으로 추론함

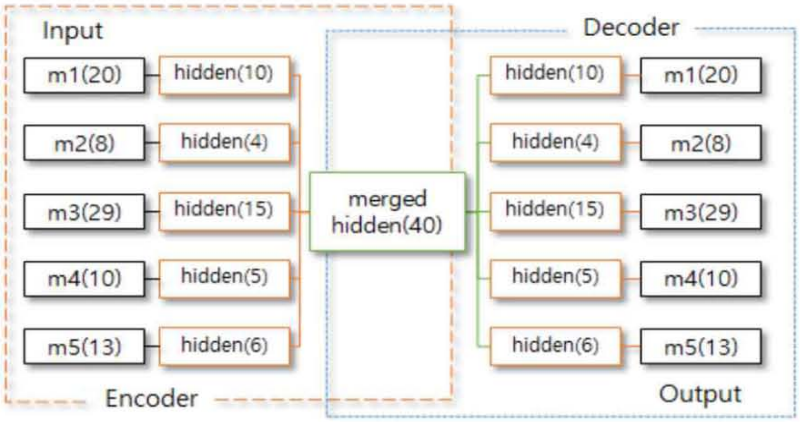
● Denoising 오토인코더 학습 결과

Model	Denoising autoencoder		Stacked denoising autoencoder	
Input Layer	80		80	
Hidden	64		64-32-64	
Output Layer	80		80	
Parameter	10,384		9,904	
Noise Factor	0.25		0.25	
Data	Training	Validation	Training	Validation
Accuracy	0.9321	0.9232	0.9287	0.9111
loss	0.4152	0.4230	0.5767	0.5868

■ **헬스 빅데이터에서 결측치 추론을 위한 멀티-모달 오토인코더**

- 기계학습에서 다양한 데이터를 활용하며, 하나의 관측대상에 대해 여러 종류의 데이터 수집하며, 데이터를 통합하여 한번에 모델링하거나 필요한 변수만을 별도로 활용하여 모델링을 진행함
- 관측 도구에 따라 객체에 대해 여러 구조의 데이터가 수집되며, 이를 하나의 트랜잭션으로 구성하는 것은 데이터의 관계나 특징이 왜곡되거나 손실될 우려가 있으며, 헬스케어는 생성된 모델의 실제 적용과정에서 주변 상황에 따라 활용 가능한 데이터가 상이할 수 있기 때문에 데이터 특성에 따라 관측대상의 모델을 여러개로 나누어서 적용하는 멀티-모달리티가 주목받고 있음
- Denoising 오토인코더를 이용하여 결측치의 추론이 가능하지만, 데이터 변수의 종류가 많아 하나의 오토인코더를 이용한 학습은 높은 계산량과 학습효율에 손실이 있음. 오토인코더는 일반적인 신경망의 한 종류로써 학습 방법이나 구성에 따라 성능의 격차가 크게 나타나며, 특히 영향력이 적은 변수의 경우 가중치가 0에 수렴할 경우 의미는 적지만 유효한 값이 남아있기 때문에 모델의 계산량을 증가시키는 원인이 됨
- 헬스케어 모델은 개인화를 위해 소형 디바이스 혹은 스마트폰을 중심으로 개발되고 있기 때문에 계산량이 적은 효율적인 구조의 모델이 필요함. 계산량에 따라 변수의 분류를 고려하여 학습데이터를 재구성하고 오토인코더를 멀티-모달로 구성하는 것이 필요함
- 80개 변수들은 만성질환 유병여부, 만성질환 의사진단여부, 만성질환 첫 진단시기, 만성질환 현재 치료여부, 신체활동여부, 운동정보, 건강검진정보, 식이영양, 스트레스 여부 등으로 구성되고, 국민건강영양조사는 hierachically 구조로 되어있으며, 항목에 따라 상위 Class와 하위 Class로 분류할 수 있음
- 상위 Class가 동일한 '고혈압 유병여부', '당뇨 유병여부', '고지혈증 유병여부', '관절염 의사진단여부', '고혈압 첫진단 시기' 등을 유병여부 클래스로 분류가 가능하며, 만성질환, 신체활동, 건강검진정보, 식이영양, 주관적 건강상태를 상위 Class로 하는 싱글-모달을 사용함
- 변수들의 분류 결과 만성질환(m1)에 20개, 신체활동(m2) 8개, 건강검진(m3) 29개, 식이영양(m4) 10개, 주관적 건강상태(m5) 13개로 5개의 모달로 구성함. 멀티-모달 Denoising 오토인코더에서 파라미터는 기존과 동일하게 하고, 분류된 입력 노드의 수에 따라 히든 노드의 구조를 변경함

● 멀티모달 Denoising 오토인코더 과정



- 각각의 싱글-모달은 Class에 분류된 변수의 숫자에 따라 입력값이 다르며, 입력보다 작은 중간 hidden 값을 가지게 되며, 인코더 부분에서 만성질환(m1)은 20×10, 신체활동(m2)는 8×4, 건강검진(m3) 29×15, 식이영양(m4) 10개×4, 주관적 건강상태(m5)는 13×6으로 구성하고, 이를 합쳐서 40개의 노드를 가지는 히든 레이어를 추가하고, 디코더는 인코더와 대칭으로 구성함
- 이머징 헬스 리스크 예측을 위해 도출해야 하는 레이어 m1(20)은 싱글-모달 만성질환(m1)이 20개의 입력노드로 구성하여 각각의 클래스에 따른 특징추출 및 학습이 가능하고, 각각의 싱글-모달을 분리하여 새로운 신경망 모델에 초깃값으로 활용 가능함

■ **상황 딥러닝 모델을 통한 이머징 리스크 상황 예측**[6]

- AnoGan 모델은 정상 데이터를 사용하여 학습을 진행하고, Latent Space(z)는 잠재적인 공간을 나타냄. 학습 시 Latent Space(z)는 Generator(G)를 통해 정상 데이터의 확률분포를 근사하며 학습을 진행하며, G를 통해 생성된 데이터는 Discriminator(D)가 실제 데이터라고 판별하도록 학습함
- 학습을 마친 모델은 G와 D의 파라미터들을 갱신하지 않고 Latent Space(z)을 랜덤 샘플링 후 생성기에 입력함. Latent Space는 정상 데이터의 특징들을 잠재하고 있기 때문에 생성기는 정상 데이터를 생성하고, 생성된 데이터와 Query 데이터를 비교하여 Residual Loss를 계산 및 Query 데이터는 정상과 비정상에 포함
- 계층간 가중치 정보전달을 통해 동적 가중치 및 경량화 알고리즘을 통해 상황 딥러닝 모델에 대한 엔트로피 경량화 프로세스를 진행함. 성능 향상을 위한 Gaussian 프로세스 기반 하이퍼파라미터 최적화를 통해 성능이 떨어지는 파라미터을 초기에 제거하여 최대한 빠르게 최적의 성능을 유도할 수 있는 값을 추출함[7]

- $L_R(z_r)$은 Residual Loss를 의미하고, $G(z_r)$은 생성된 데이터이며 x는 Normal Image를 나타냄. 생성기가 충분히 학습되었다면 Latent Space(z)로 정상데이터의 확률분포가 근사되고 x가 G에 들어왔을 때 x의 수치는 유사하기 때문에 Residual Loss는 0에 가까운 값이 추출됨

• Residual Loss 산출식

$$L_R(z_\gamma) = \sum |x - G(z_\gamma)|$$

- 판별기는 생성된 데이터의 확률 분포에 따라 True와 False의 여부를 결정하고, Discrimination Loss를 계산하며, z에 의해 생성되는 데이터가 생성기에 적절하게 매핑되도록 하는 Loss를 나타냄
- f(x)는 Query 데이터의 확률 분포를 나타내며, $f(G(z_r))$는 생성된 데이터의 확률 분포를 나타냄

• Discrimination Loss 산출식

$$L_D(z_\gamma) = \sum |f(x) - f(G(z_\gamma))|$$

- Loss Function은 Anomaly Score로 사용하고, 이상치 판단의 기준으로 사용하기 위해 Total Loss 계산함. Residual Loss의 값과 Discrimination Loss를 합한 값이며, $L_R(z_r)$는 Residual Loss, $L_D(z_r)$는 Discrimination Loss를 나타내고, λ는 반복적으로 진행한 성능평가 실험에서 도출한 결과인 가장 효율적 성능을 보이는 0.1로 진행하여 Total Loss는 Anomaly Score로 사용함. 입력으로 들어오는 Query 데이터가 정상 또는 비정상으로 분류되는 판단의 기준이 됨

• Total Loss

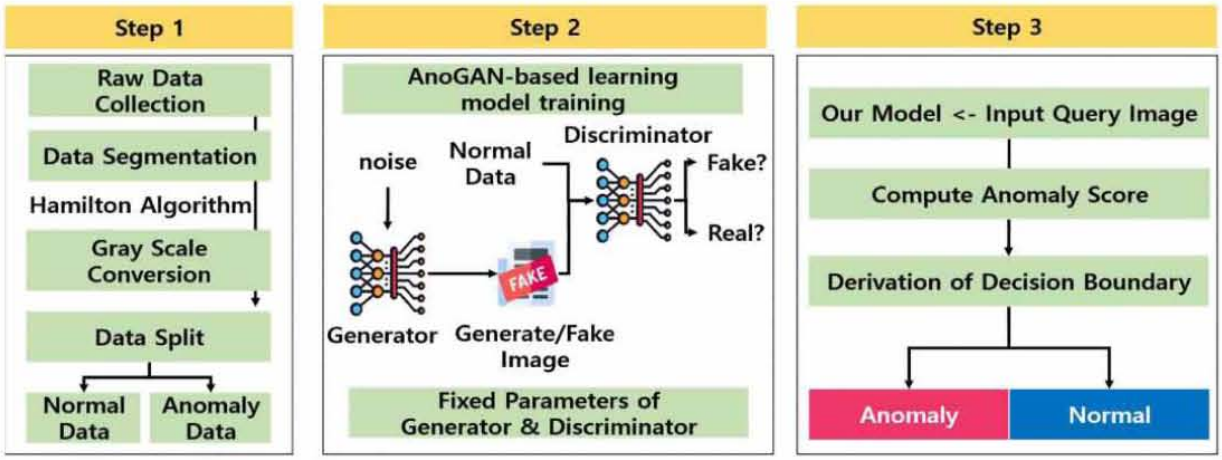
$$Total\,Loss = (1-\lambda) \cdot L_R(z_\gamma) + \lambda \cdot L_D(z_\gamma)$$

6) J. W. Baek, D. H. Shin, K. Chung, Graph Neural Network based Recommendation Model for Traffic Safety, The International Conference on Convergence Technology 2021, pp.273-274, 2021.

7) 김민정, 전병욱, 권혜정, 백지원, 정경용, BERT Bi-GRU 성능 향상을 위한 Gaussian 프로세스 기반 하이퍼파라미터 최적화, 2021 한국인터넷정보학회 추계학술발표대회, pp. 155-156, 2021.

■ 심전도 데이터에서 개선된 AnoGAN을 이용한 의사결정 경계 기반 이상치 탐지 모델[8)]

• 이상치 탐지 모델 프로세스 구성도



- ECG 데이터 수집 및 전처리 단계에서 데이터는 Physionet에서 제공하는 MIT-BIH 부정맥 ECG 데이터를 수집하고, 전처리는 데이터에 포함되어 있는 측정시 생기는 잡음을 제거하기 위해 필터링을 함. 또한 모델의 입력을 심전도의 비트별로 구성하기 위해 세그먼테이션을 하여 심전도 데이터의 R-pear 점을 기준으로 하여 세그먼테이션 설정함
- 비트별로 구성된 데이터는 모델의 채널과 맞춰 주기 위해 3차원의 RGB 색의 채널을 2차원인 그레이 스케일의 채널로 변환하고, 2D-Convolution Layer를 사용하기 때문에 데이터를 이미지 형식으로 저장함[9)]
- 이상치 탐지를 위한 개선된 AnoGAN 기반 이상치 탐지 모델 설계, 훈련 및 검증 단계에서는 전처리 과정을 거친 Normal Image를 사용하여 학습을 진행하고, 분류를 위한 Anomaly Score의 변화를 통해 의사결정 경계를 결정함

▪ 개선된 AnoGAN 기반 학습을 위한 부정맥 데이터 수집 및 전처리

- 모델의 학습을 위해 Physionet에서 제공하는 MIT-BIH 부정맥 ECG 데이터를 사용하며, Boston's Beth Israel Hospital에서 수집된 데이터이고, 32~89세의 남자 25명, 23~89세 여성 22명이며 총 47명을 대상으로 수집된 30분가량의 ECG 데이터 48개로 구성함

• Physionet MIT-BIH에서 사용자의 특성별로 구성된 데이터

Record Series	No. of MIT-BIH Records
100	100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 111, 112, 113, 114, 115, 116, 117, 118, 119, 121, 122, 123, 124
200	200, 201, 202, 203, 205, 207, 208, 209, 210, 212, 213, 214, 215, 217, 219, 220, 221, 222, 223, 228, 230, 231, 232, 233, 234

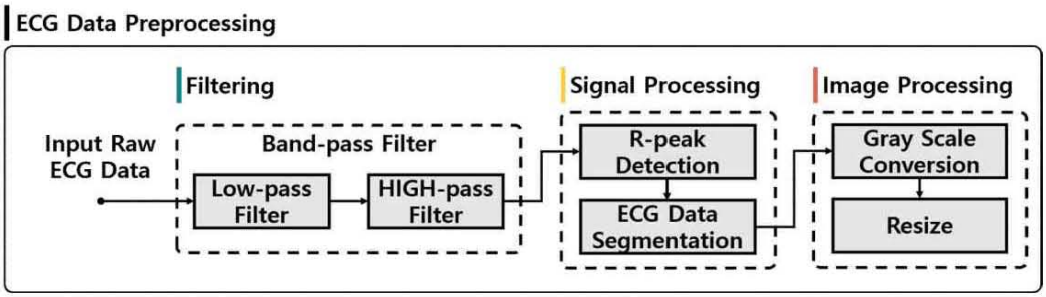
- 23개의 기록(100 series)은 외래환자(40%)와 입원환자(60%) 4,000명 중 임의로 선택되었으며, 나머지 25개의 기록(200 series)은 의학적으로 중요하거나 특이한 증상을 가진 환자들로 구성됨
- 해당 데이터는 심전도 기록으로 구성된 신호 데이터이며, 기록은 2채널로 구성되며 대부분 사지유도 방식

8) D. H. Shin, R. C. Park, K. Chung (2020) Decision Boundary-Based Anomaly Detection Model using Improved AnoGAN From ECG Data, IEEE Access, 8, 108664-108674.
9) J. W. Baek, S. E. Ryu, B. K. Koo, S. S. Park, K. Chung, Risk Prediction using Correlation from Traffic Modality, The International Conference on Convergence Technology 2020, pp.110-111, 2020.

(MLII)과 V1방식(예외적으로 V2, V4, V5로도 기록된 경우도 있음)으로 이루어져 있음

- 데이터는 0.1~100Hz의 대역 필터링과 360Hz의 샘플링 속도로 균일하게 구성되며 15가지 클래스의 정상 리듬이 포함된 부정맥 데이터이며, 심전도 전문가로부터 데이터의 정보가 주석처리 되어있음. 수집된 데이터를 모델의 입력 데이터로 사용하기 위해 전처리 진행함

● 심전도 데이터 전처리 과정



- Raw ECG Data는 피실험자를 통해 실제로 측정된 데이터이며, 원래의 데이터를 그대로 사용하게 되면 측정 시 발생하는 여러 잡음으로 인해 ECG Data를 통한 정확한 학습이 어렵기 때문에 ECG Data에 포함되어 있는 측정시 발생한 잡음을 제거하는 필터링 과정이 필요함. 필터링을 위해 Hamilton Algorithm(Low Pass Filter(LPF) 및 High Pass Filter(HPF))을 사용함
- LPF는 ECG Data의 심전도 파형에서 전원의 간섭으로 생기는 잡음을 제거하기 위해 사용하고, 심전도 베이스라인의 변동을 감소시키기 위해 HPF를 사용함
- 필터링된 이미지는 비트별로 구성하기 위해 R-peak점을 탐지하고, 탐지된 R-peak 점을 기준으로 전/후에 따라 P파와 T파가 존재하며, 이를 비트로 구성하기 위해, R-peak 점을 기준으로 0.3초 전과 0.4초 후의 범위에서 세그먼트를 진행함

- 비트별로 구성된 데이터는 모델에서 2D-Convolution Layer를 사용하기 때문에 신호 데이터를 이미지로 변환하고, 변환된 이미지는 R, G, B로 구성된 실수로 3차원으로 구성함
- 모델의 채널은 이미지의 채널과 동일함. 모델의 채널수가 증가하게 되면 필터수도 증가하게 되어 연산의 시간이 증가하기 때문에 연산시간을 줄이기 위해 Gray Scale Conversion을 통한 2차원으로 채널 축소함
- 모델의 입력크기(64×64)와 이미지의 크기를 맞추기 위해 resize를 하고, 전처리된 이미지는 모델의 학습과 검증에 사용하기 위해 Association for the Advancement of Medical Instrumentation(AAMI) 권고에 102, 104, 107 및 218레코드는 좋지 않은 신호 품질이기 때문에 제거하고 데이터를 재구성 진행함[10)]

- MIT-BIH 데이터의 15개 클래스를 AAMI의 권고에 따라 5개 클래스로 재구성하여, AAMI의 Heart Beat 클래스는 Normal(N), Supraventricular(S), Ventricular(V), Fusion(F), Unknown(Q) 신호로 구성함. 분류된 데이터로 모델을 통해 학습을 진행. 정상 데이터만을 사용하여 모델의 학습을 구성함으로 N Class를 사용하고, N Class 이외의 클래스(S, V, F, Q Class)는 검증데이터로 사용함

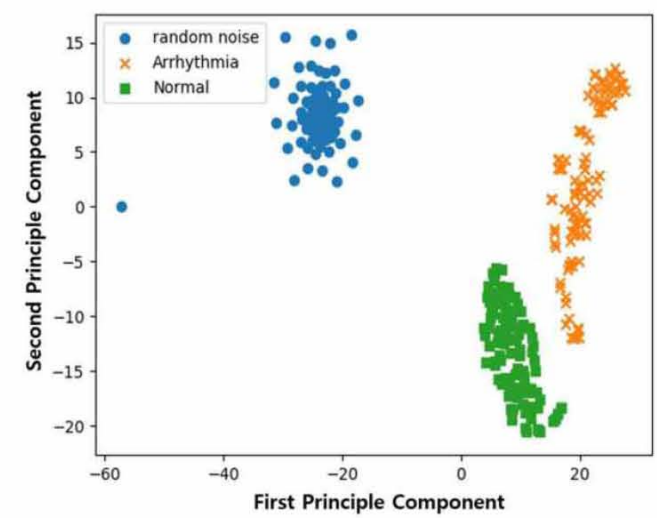
- 데이터의 올바른 분류를 확인하기 위해 데이터의 특징 추출하여 시각화하며, 기존의 고차원 데이터의 차원 축소를 위해 PCA(Principal Components Analysis)를 사용한 선형 차원 축소 방법은 중첩 현상이 발생한다는 문제점이 존재 하기 때문에 t-분포 확률적 임베딩(t-SNE)을 사용함으로써 차원 축소 및 시각화를 하는 비선형적 방법으로 고차원상의 군집간의 관계를 저차원상에서도 특징을 표현 가능함
- t-SNE는 두 가지의 주요 단계로 진행되며, 첫 번째 단계는 특징 추출을 통해서 객체간의 유사도에 따라 확률 분포를 구성하고, 두 번째 단계는 저차원 공간에서 객체에 대한 확률분포를 정의하여, 객체의 위치에 관한 두 분포사이의 Kullback-Leibler divergence (KL divergence)를 최소화 프로세스를 통해 진행함

10) 권혜정, 전병욱, 김민정, 백지원, 정경용, 도로교통 이상 탐지를 위한 어텐션 마스킹 기반 적대적 학습, 2021 한국인공지능융합기술학회 추계학술대회, pp. 21-22, 2021.

• MIT-BIH의 Heart Beat Class와 AAMI의 Heart Beat Class 구성

MIT-BIH Heartbeat Class	AAMI Heartbeat Class
Normal (N) Left Bundle Branch Block (L) Right Bundle Branch Block (R)	Normal (N)
Atrial Premature (A) Aberrated atrial Premature (a) Nodal (junctional) Premature (J) Supraventricular Premature (S) Atrial Escape (e) Nodal Escape (j)	Supraventricular(S)
Premature Ventricular Contraction (V) Ventricular Escape (E)	Ventricular(V)
Fusion of Ventricular and Normal (F)	Fusion(F)
Paced (/) Fusion of Paced and Normal (f) unclassified (Q)	Unknown(Q)

• t-SNE 기반 차원축소를 이용한 데이터 시각화 결과

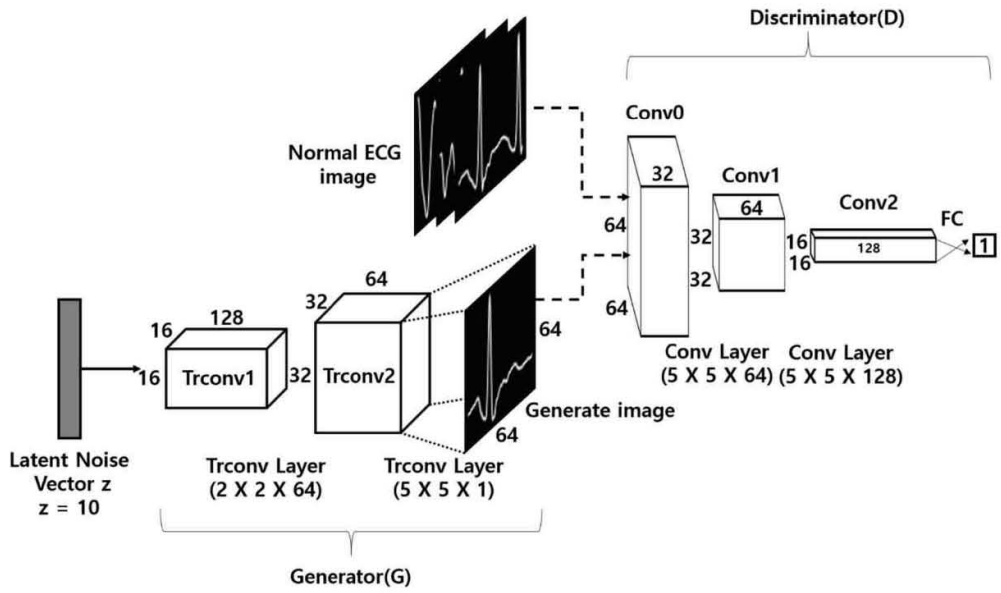


- ●은 Random noise, ×는 Arrhythmia, ■은 Normal을 나타내고, Random noise 데이터와는 달리 Arrhythmia와 Normal 데이터는 각각 정상과 부정맥이라는 다른 범주이지만 같은 심전도 데이터이기 때문에 근접한 위치에서 군집화하여, t-SNE의 장점인 전역(global) 구조와 지역(local) 구조를 모두 유지하고 있음을 나타냄
- 전역 구조와 지역 구조를 모두 유지한다는 것은 군집화가 적절하게 생성되었을 뿐만 아니라 데이터간의 특징이 비슷한 경우 위치적으로 가까운 공간에 군집화가 이루어짐을 확인함

■ **개선된 AnoGan기반 이상치 탐지 모델 설계, 훈련 및 검증 단계**

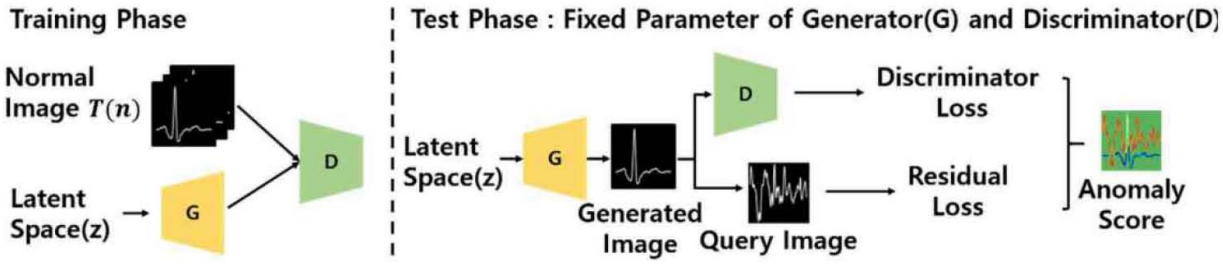
- 분석에서 예측을 위한 모델링은 지도학습 모델을 사용하여도 이상치 탐지는 가능하지만 헬스 분야에서는 병변을 가지는 비정상 데이터보다 병변을 가지지 않는 정상 데이터가 현저히 적어 데이터 수집에 어려움이 있기 때문에 지도학습 모델은 데이터의 불균형과 불충분하다는 문제점이 있음
- 부정맥 데이터를 이상치로 정의하고, 정상 데이터만을 사용하여 학습하는 GAN 기반의 비지도 학습 모델인 AnoGAN을 사용

● 이상치 탐지 모델의 네트워크 구성



- GAN 기반 AnoGAN을 사용하여 구성되며 주요 구성은 Convolution Layer와 Transposed Convolution Layer로 구성함. 활성화 함수는 Reaky Relu를 사용하며 Dropout를 이용하여 학습시 과적합을 방지하고, 입력 값은 정상 데이터를 이용하며 크기는 64 × 64로 사용함. 대상 도메인으로부터 최적화한 이상치 탐지 네트워크를 통해 부정맥을 탐지하며, 이상치 탐지 모델에서 학습을 하기 위해 훈련단계와 검증단계로 구분하여 진행함

● 이상치 탐지 네트워크에서 훈련 및 검증 단계



- 훈련단계에서 비트별로 구성한 정상 이미지를 사용하며, n은 학습 이미지의 수를 나타내고, Generator(G)와 Discriminator(D)를 학습시키고 정상 이미지의 확률 분포 정보를 샘플링된 저차원 노이즈인 z를 통해 Latent Space에 근사함
- GAN의 특성인 G와 D의 적대적 학습 후에는 G로부터 Latent Space(z)를 통해 이미지 생성함. 이는 훈련된 생성기와 정상 이미지의 분포 정보를 가지는 Latent Space(z)를 통해 검증에 사용하고, 검증 단계에서는 G와 D의 파라미터를 갱신하지 않음
- Query Image(x)가 주어지면 이상치 탐지 모델을 통해 x와 유사한 이미지 G(z)를 만드는 최적의 z를 찾는 것을 목표로 진행하며, 최적의 z를 기반으로 이미지를 생성하고 x와 비교하여 손실함수를 통해 이상치 스코어(Anomaly Score)를 계산함

- 두 가지 손실 함수를 정의하며, 첫 번째로 G에서 생성된 이미지 G(Z)와 Query Image(x)의 차이를 비교하는 Residual Loss(R Loss)이고, 두 번째는 G(Z)가 만드는 이미지가 D로 입력될 때 데이터의 확률 분포가 True 또는 False로 판단하는 Discriminator Loss(D Loss)가 있음
- R Loss와 D Loss는 동시에 학습을 진행해도 R Loss는 줄어들지 않고 D Loss만 낮아지는 현상이 발생하기 때문에 학습의 안정화를 위해 반복적인 실험을 하여 Generator의 학습 횟수는 4번 Discriminator의 학습은 1번으로 부여하는 것이 가장 적합하게 Loss의 균형을 유지한다는 것을 보임

● 학습 안정화를 위한 Loss 균형함수

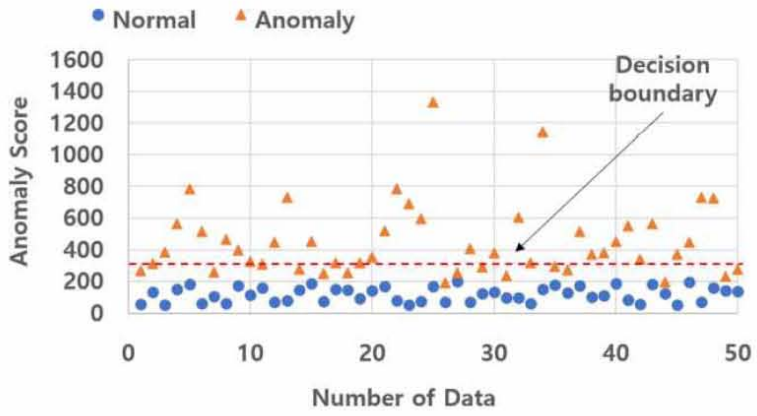
Function: Loss Balance Function for Learning Stabilization.

Input: Train Data(x)
Output: G and D Loss
Train_Phase(x)
  D ← Discriminator_model
  G ← Generator_modela
  for epoch(n)
    N ← Random Noise
    Real_Image ← Load data
    Generated_image ← G. predict(N)
    X ← Concatenate (Image, Generated_image)
    D.train(X)
    D.trainable ← False
    for range (3):
      G. train(N)
    D.trainable = True
return D, G

■ **분류를 위한 이상치 스코어를 통한 결정 경계 산출**

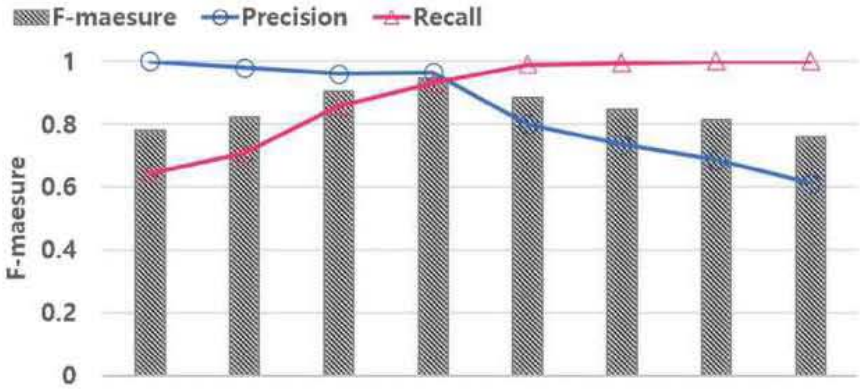
- 정상과 부정맥 분류를 위해서는 학습을 마치고 가중치가 고정된 모델에서 query 이미지를 입력하여 이미지와 생성기를 통해 생성된 이미지의 비교를 수행하여 이상치 스코어를 산출함
- 기존의 AnoGAN 모델의 단점은 결정 경계가 주관적으로 정해야 하는 문제점이 있음. 결정 경계에서 두 개의 클래스를 가지고 있는 통계적인 분류에서 벡터 공간을 각각의 클래스에 따라 두 개의 집합으로 나누는 초평면이고, 분류기는 결정 경계의 한쪽에 있는 모든 점을 한 클래스, 다른 한쪽에 있는 모든 점을 다른 클래스에 속하는 것으로 분류함
- 이머징 리스크에 대한 판단을 위해 이상치 스코어로 결정 경계를 정의하고, 이를 통해 정상과 부정맥을 분류하며, 결정 경계를 350으로 주관적으로 결정하게 되면 부정맥인 데이터가 정상으로 분류함

● 이상치 스코어를 통해 결정된 주관적인 결정 경계 판단



- 정밀도와 재현율의 곡선이 교차하는 지점은 이상치 스코어가 200인 경우이며, 200 이상인 경우에는 정밀도는 낮아지지만 재현율은 높아짐. 결정 경계가 높아지면서 정상으로 판단하는 범위가 넓어져서 재현율이 높아지며 부정맥으로 판단해야 하는 데이터도 정상으로 판단하여 정밀도가 떨어지기 때문에 반복적인 실험을 통한 정상/부정맥 분류를 위한 결정 경계는 200으로 부여하는 것이 정밀도와 재현율이 가장 높으므로 분류의 정확도를 높이는데 가장 효과적임

● 결정 경계의 변화에 따른 정밀도와 재현율



■ 교통 영상 빅데이터 처리를 위한 Yolo 기반 광원 객체 탐지[11]

- 객체 탐지는 영상 내의 물체를 식별하는 컴퓨터 비전 기술이며, 딥러닝을 이용한 객체 탐지 알고리즘은 1-stage Detector와 2-stage Detector가 있음
- Object detection 모델 중 하나로 Yolo를 1단계 검출기로 사용함. 전체이미지를 동일 크기의 그리드로 나누고 다양한 객체에 대한 bounding box를 찾아내는 알고리즘을 사용함. 이는 찾은 물체의 중심점을 기반으로 bounding box의 개수를 예측하여 신뢰도를 계산함. 신뢰도에 따라 bounding box는 많이 나타나므로 진하게 나타나게 되므로 진한 box만을 남겨 객체를 탐지함
- R-CNN(Region Based Convolutional Neural Network)은 Yolo와 다른 2단계 검출기를 사용함. 검출기와 localization의 순서로 진행되며 이는 객체가 있을 후보영역을 찾는 것과 특징추출의 과정으로 이루어짐
- R-CNN 모델은 Region proposal을 통해 후보 영역을 선택적 검색 알고리즘을 통해 파악하며 이 과정에서 색, 질감, 경계선의 요소가 사용됨. 이 알고리즘에서 추출된 이미지를 컨볼루션 레이어의 입력값으로 사용되므로 1단계 검출기 모델에 비해 복잡도가 높아 연산량 및 수행시간이 더 소요됨. 또한 역전파의 수행이 어려우므로 이를 발전시킨 R-CNN의 변형 모델인 Fast-R-CNN 등의 모델이 다양하게 등장함

▪ 데이터 수집 및 전처리

- BDD100k Tracjubg Challenge에서 공개된 데이터 세트를 사용함. 이는 미국의 다양한 지역에서 자동차를 촬영한 영상이며 날씨, 위치 및 시간대로 데이터가 구성됨. 총 10만 개의 영상으로 1280x720의 크기를 가짐
- 밤 시간대 도로의 광원객체 탐지 목적을 위해 해당 데이터 세트의 nighttime 영상만을 사용하며 특징 추출이 어려운 흔들림 사진 또는 객체가 존재하지 않는 사진을 제외하여 사용함

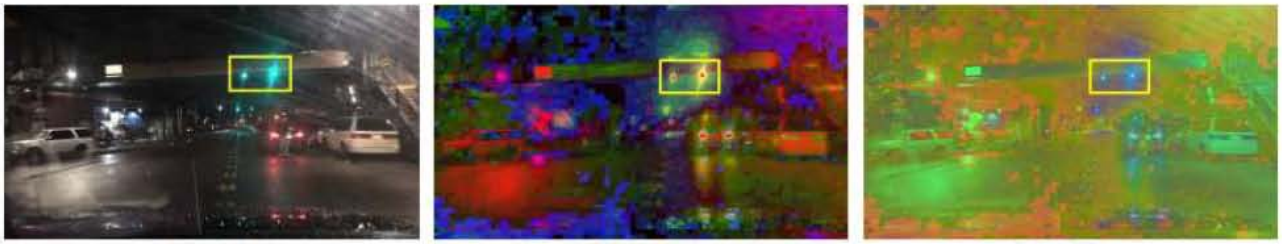
▪ 색상 모델의 변환을 이용한 이미지 처리[12]

- 광원객체 탐지를 위해 전처리된 이미지가 요구됨. 두 가지 색상 모델 HSV와 YCbCr을 이용하여 딥러닝 학습에 적합한 입력 이미지로의 처리를 진행함
- 일정 범위의 순수 색 정보를 가진 H를 이용한 HSV 모델은 RGB 모델에 비해 색의 분류가 용이한 특성이 있음. 이러한 특징을 이용해 HSV 모델을 통해 효과적인 색 분류 및 YCbCr의 밝기와 색차 성분 분석을 통해 야간도로에서의 광원객체 탐지 및 인식률을 높임

- YCbCr은 색 공간으로, 주로 영상 시스템에서 사용되며, Y는 밝기 성분 값이고, Cb와 Cr은 색차 성분이며, 광원의 빛번짐 현상은 주변에서 노출되는 빛의 양과 밀접한 관련이 있고, 광원의 단위 입체각으로 발산하는 빛 성분을 고려하는 YCbCr 모델을 통해 빛 번짐 현상의 보완이 가능함. 따라서 HSV 색상 모델을 통해 색 분류 과정을 거친 후 YCbCr을 통해 빛번짐 현상을 보정함으로써 특징 추출이 용이한 영상으로 변환함

11) 강지수, 심세은, 조선문, 정경용 (2020) 교통 영상 빅데이터 처리를 위한 Yolo 기반 광원 객체 탐지, 융합정보논문지, 10(8), 40-46.
12) 백지원, 김민정, 권혜정, 강지수, 정경용, YOLO를 이용한 객체 지식 그래프 기반 이상 탐지, 2021 한국인공지능융합기술학회 하계학술대회, pp. 27-28, 2021.

● 두 단계 색상 모델을 통한 전처리 과정



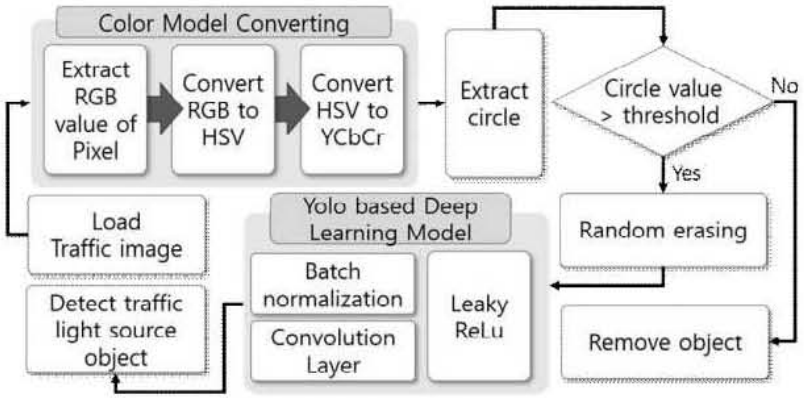
(a) Raw Data (b) HSV (c) YCbCr

- (a)는 원본 데이터로 특징 추출이 어려워 입력 데이터로 적합하지 않음
- (b)는 HSV 색상 모델로 전처리한 이미지이며, 보다 더 단순화된 색상 값과 광원 특징이 잘 나타남
- (c)는 HSV를 통해 처리된 데이터에서 YCbCr 색상으로 변환한 이미지이며, (b)보다 빛 번짐 현상이 줄어들며 광원의 밝기와 색차 성분을 분석하여 이미지 처리함

■ 도로 상황에서 딥러닝 기반 광원 탐지

- 신호등 이미지의 경우 일반적으로 신호등은 전광판, 표시판, 가로등에 의해서 일부분이 가려지는 현상이 발생하여 가려진 이미지 데이터에서 객체를 탐지하는 학습을 진행하기 위해 이미지를 자르는 Cropping 방식을 사용함
- 원본 이미지에서 의도적으로 일부분을 가림으로써 물체의 전체적인 구조를 보존하도록 하는 Random Erasing 방법을 사용하여 임의로 지정된 영역의 픽셀에 값을 다시 할당하고, 다른 물체에 의해 가려진 영역을 통해 원본 이미지에 노이즈를 주어 모델이 과적합되는 문제를 해결함

● 교통 영상 빅데이터 처리를 위한 Yolo 기반 광원 객체 탐지의 흐름도



- Yolo의 CSPDarknet53 모델은 다른 모델보다 파라미터 수가 많음에도 성능이 우수하게 나타나며, 큰 물체뿐만 아니라 작은 물체마저 탐지할 수 있기 때문에 실시간적으로 신호등을 탐지하고 신호등의 색상을 분류함
- CSPDarknet53 모델을 통해서 신호등의 탐지와 색상 분류 태스크를 수행하고, LeakyReLU를 활성화 함수로 사용하여 loss를 감소시킴
- 원본 교통 도로 영상 데이터에서 RGB 픽셀값을 추출함으로써 RGB 색상 모델을 HSV 색상 모델로 변환함. 변환된 교통 도로 이미지를 다시 YCbCr 색상 모델로 변환함으로써 특징을 추출하는 작업을 원활하게 하도록 하며, YCbCr 색상 모델을 적용한 학습 데이터에서 원래의 객체를 추출함으로써 원의 크기를 산출하여 일정한 임계값을 넘는지 확인하는 과정을 진행함
- 지정한 임계값을 넘지 못하는 객체 이미지는 신호등 후보군에서 제외되며, 이와 같은 과정에서 신호등과 비슷한 특징을 가지고 있는 광원 객체를 제외하고 신호등만 탐지함. 특정 객체가 임계값을 넘게되면 Random Erasing 방식을 통해서 Yolo를 학습시키고, 작은 객체마저 탐지할 수 있는 CSPDarknet53 모델을 적용함. 활성화 함수로 LeakyReLU를 사용하며 배치 정규화와 합성곱 레이어를 통해 광원 객체를 탐지함

■ TF-IDF 기반 순위 스코어를 이용한 연관지식 그래프의 최적화[13)]

- 연관규칙은 어떠한 아이템의 집합이 얼마나 빈번히 발생하는지에 대한 규칙을 알아내는 알고리즘이며, 트랜잭션에 있는 데이터를 기반으로 A라는 아이템을 구매한 고객에 대해서 B라는 아이템을 구매할 확률 발견하는 비즈니스 타겟 마케팅에서 사용되는 방법임
- 이머징 헬스 예측의 헬스케어 분야에서는 만성질환간의 연관규칙, 건강상태에 따른 질병과의 연관규칙에 따라 정보를 획득할 수 있으며, 연관규칙 알고리즘을 적용하기 위해 데이터 세트를 트랜잭션으로 구성

• 트랜잭션은 Transaction ID와 Itemset

Transaction ID	Itemset
1	A, B, C, D
2	B, C, E
3	A, B, E
4	B, C, D, E
5	A, B, D, E
...	...

- 1번 Transaction ID를 가진 고객이 마켓에서 A, B, C, D의 Item을 구매한 것으로 해석할 수 있음. 트랜잭션 데이터를 이용하여 연관규칙을 표현할 때, A라는 아이템을 구매한 고객이 B라는 아이템을 구매할 경우에 대한 연관규칙이 생성됨
- Lhs는 Left-hands-side를 나타내며, Rhs는 Right-hands-side를 나타내고, Lhs => Rhs는 왼쪽 항목과 오른쪽 항목이 연관이 있음을 나타냄

• 연관규칙 데이터

Lhs => Rhs	Support	Confidence	Count
A => B	3/5	3/3	3
B => A	3/5	3/5	3
B => C	3/5	3/5	3
C => B	3/5	3/3	3
...	...	...	...

- 연관규칙에서 아이템간의 규칙을 알아내기 위해 사용되는 척도로 지지도(support), 신뢰도(confidence), 향상도(lift)를 사용함
- 교통사고 관련 코퍼스에서 정보를 추출하기 위해 사용하는 교통사고 관련 텍스트는 국토교통부에서 제공하는 보도자료의 교통사고 관련 섹션에서 텍스트 문서를 수집함. 선택한 교통사고 도메인에서 각 보도자료를 전처리하여 단어를 추출하여 트랜잭션을 생성하고, 해당 보도자료 트랜잭션을 바탕으로 텍스트 마이닝을 적용할 수 있음

• 텍스트 마이닝 도메인의 국토교통부 보도자료를 바탕으로 구성한 트랜잭션 데이터와 연관규칙 데이터

Transaction ID	Itemset
1	accident, traffic, driver, drunk
2	traffic, safety, decrease
3	road, elderly, driver, accident
4	police, crackdown, drunk, driver
5	student, bus, inspect, accident
...	...

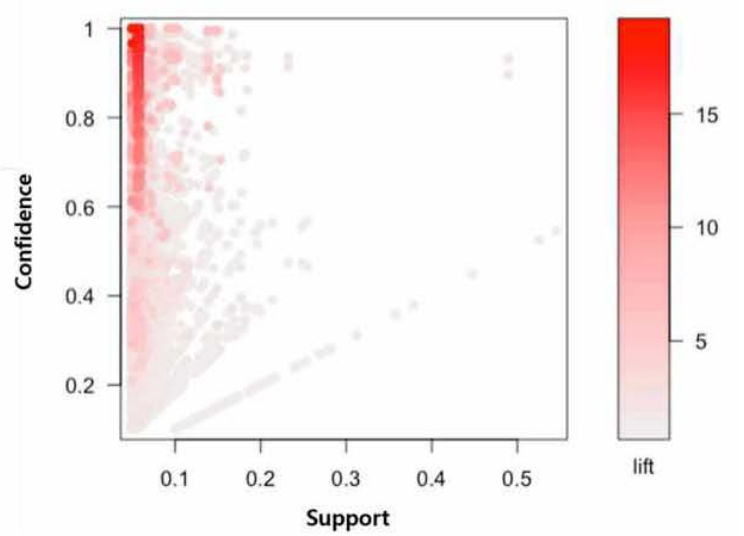


Lhs => Rhs	Support	Confidence	Count
accident => traffic	20/791	9/791	20
traffic => accident	20/791	6/791	20
driver => accident	54/791	42/791	54
accident => driver	54/791	21/791	54
...	...	...	...
crackdown, drunk => accident	9/791	7/791	9
accident => crackdown, drunk	9/791	3/791	9

13) H. J. Kim, J. W. Baek, K. Chung (2020) Optimization of Associative Knowledge Graph using TF-IDF based Ranking Score, Applied Science, 10(13), 4590-4910.

- 1번 뉴스에는 accident, traffic, vehicle, drunk의 추출된 단어가 존재하고 accident와 같은 단어가 등장했을 때, traffic 단어가 등장하는 경우 등에 대한 연관규칙들이 생성됨

• 스트림 데이터를 이용하여 연관규칙 생성 결과



- 연관규칙 결과를 이용하여 지식 그래프를 생성할 수 있으며, 지식그래프는 서로 관련 있는 정보와 지식을 연결하여 표현한 것을 의미하고, 잠재적인 정보와 지식을 추론할 수 있음. 연관규칙의 결과로 지식그래프를 생성하게 되면 서로 관련된 규칙의 정보와 지식을 직관적으로 파악하고, 유의미한 정보 및 지식을 추론할 수 있음
- 가로축은 지지도를 나타내며 세로축은 신뢰도를 나타내고, 향상도는 색이 진할수록 높아지며 이는 연관관계가 긍정적임을 의미함
- TF-IDF(Term Frequency - Inverse Document Frequency)는 정보검색과 텍스트 마이닝에서 단어에 대한 가중치를 계산하는 방법이며, 단어가 문서에서 어느정도 중요도를 가지는지 판단하기 위해 사용됨

• TF-IDF(Term Frequency - Inverse Document Frequency)

$$TF-IDF(t,\ d,\ D) = tf(t,\ d) * idf(t,\ D)$$

- TF는 term frequency로서 문서의 특정한 단어의 빈도수를 나타내며, 하나의 문서에 특정 단어가 몇 번 등장했는지를 의미함. TF의 값은 특정 단어의 등장 횟수가 많아질수록 무한히 발산하는 특성을 보임
- IDF(Inverse Document frequency)는 다수의 문서에 공통적으로 등장하는 단어에 역수를 취한 값을 의미함
- t는 특정 단어가 포함된 문서의 개수를 뜻하고 D는 전체 문서의 개수를 뜻하며, DF는 문서에서 특정 단어의 빈도수를 표현하기 때문에 t를 D로 나눔

• DF(Document Frequency) 산출식

$$DF(t,\ D) = \frac{t(n)}{D(n)}$$

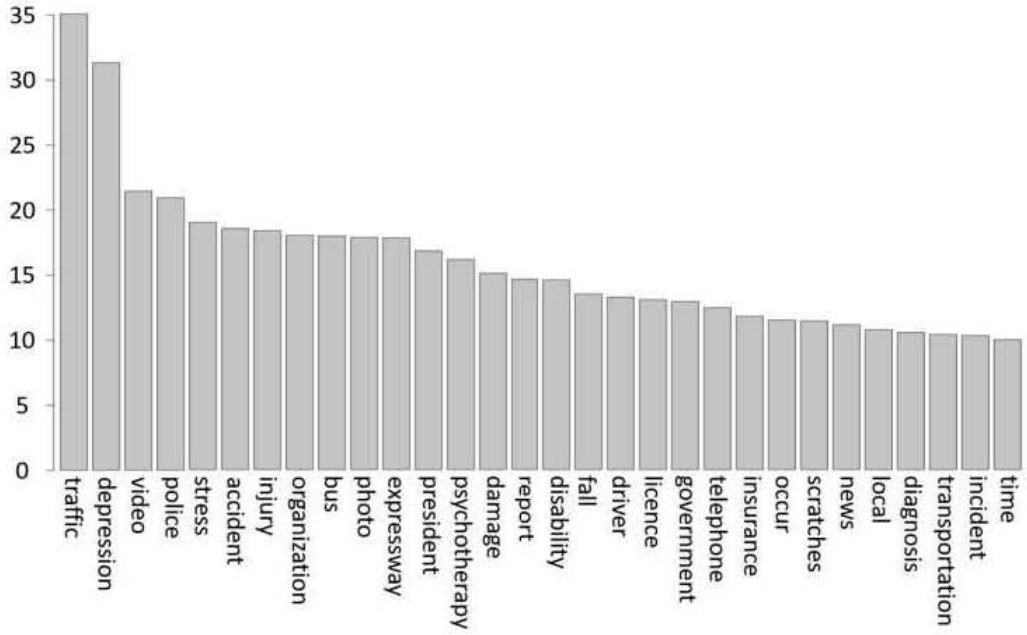
- df값을 계산하는 이유는 전체 문서에서 특정한 단어가 등장한 문서가 많을수록 단어 자체가 변별성이 부족하다고 판단되기 때문이며, 흔하게 나타나는 단어('a', 'an', 'the' 등)에 대한 가중치를 낮춰주기 위해서 df 값을 계산함
- df의 값이 높을수록 전체 결과값이 낮아져야 변별성이 없는 단어에 대한 가중치를 낮춰줄 수 있기 때문에 df를 역수로 취하여 idf로 나타내지만 idf값을 취함으로써 특정 단어가 등장하지 않을수록 기하급수적으로 값이 커지는 경향을 보이는 문제점을 해결하기 위하여 df의 역수값에 log를 취함

• log를 적용한 idf를 계산식

$$IDF(t,\ D) = \log\left(\frac{D(n)}{t(n)}\right)$$

- 문서에서 단어 t의 빈도수 TF와 전체 문서에서 단어 t가 나타나는 빈도수를 역수로 취한 값 IDF값을 곱하여 TF-IDF 값을 도출하게 됨. 특정 문서에서 특정한 단어의 빈도수가 높고, 전체 문서에서는 해당 단어를 포함한 문서가 적으면 TF-IDF의 값이 높아지게 됨. 모든 문서에서 등장하는 단어들에 대한 가중치를 낮춰주어 특정 문서에만 등장하는 중요도가 높은 단어에 가중치를 높게 하여 중요한 단어를 선별함
- MBC(Munhwa Broadcasting Corporation)에서 제공하고 있는 뉴스 스트림 데이터에서 1,700개의 교통사고 토픽 관련 텍스트 데이터를 수집하였고 코퍼스로 변환함
- 수집한 문서의 단어에 대한 TF-IDF 가중치를 계산한 결과를 바탕으로 전체 문서에 대한 단어의 가중치를 합산하여, 가중치가 높은 순으로 내림차순으로 정렬하여 TF-IDF Rank Table을 생성함

• 뉴스 스트림 데이터에서 TF-IDF를 이용한 단어의 가중치 측정 결과

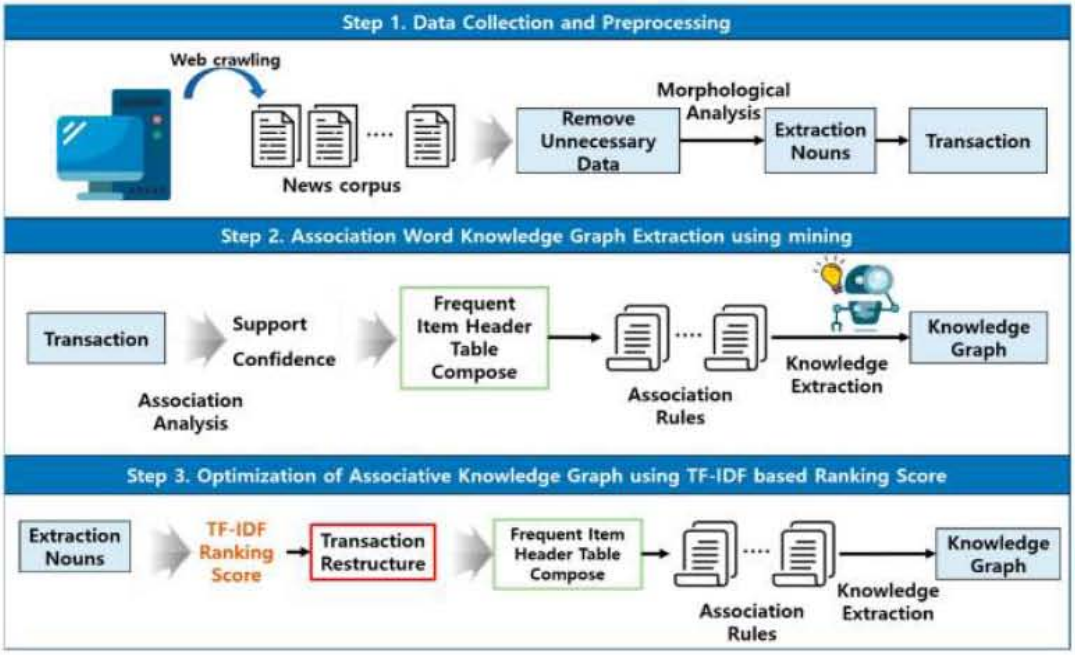


- Traffic 단어가 전체 문서에 대해 TF-IDF 가중치 점수가 35로 가장 높게 측정되고, time은 가중치 점수가 약 10점으로 가장 낮게 측정됨. TF-IDF Rank Table의 순위를 바탕으로 효율적인 연관규칙 발견을 위해 트랜잭션을 재구성하며, 교통사고 토픽 텍스트에서 추출된 단어의 상위 20% 이하의 TF-IDF Rank가 부여된 단어를 삭제하는 과정을 진행함
- 실시간으로 생성되는 다량의 데이터에서 유의미한 정보를 파악하기 위해 연관규칙 알고리즘의 속도와 효용성 개선이 필요함

- 기존의 연관규칙을 통해 생성된 지식 그래프에서 중요도가 낮은 단어들의 정보들이 포함되어 있어 정보제공의 효율성이 떨어지는 문제점을 해결하기 위해 TF-IDF 기반 순위 스코어를 통해 중요도가 낮은 단어들은 제거하여 지식 그래프 생성함
- 첫 번째 전처리 단계에서는 크롤링을 통해 교통사고 토픽 관련 뉴스와 교통안전 토픽 관련의 뉴스를 실시간으로 수집하여 코퍼스 구성하기 위해 뉴스 코퍼스에서 불필요한 데이터를 제거하고 형태소분석을 통해 단어를 추출하여 트랜잭션을 생성함
- 두 번째 단계는 마이닝을 이용한 연관단어 지식 그래프를 추출. 트랜잭션에서 지지도와 신뢰도를 통해 frequent item header table을 생성하고 연관규칙을 발견하여 지식을 추출하고 그래프를 생성함
- 마지막 단계에서는 TF-IDF를 이용하여 첫 번째 단계에서 추출한 단어들의 순위를 결정함. 순위가 낮은 단어들은 중요성이 낮다고 판단하여 제거하며 트랜잭션을 재구성함
- 연관규칙을 통해 유의미한 지식을 추출하고 그래프를 생성하고, 생성된 지식 그래프에서는 교통사고, 교통

안전과 관련된 지식베이스 구성과 이머징 리스크 예측이 가능함

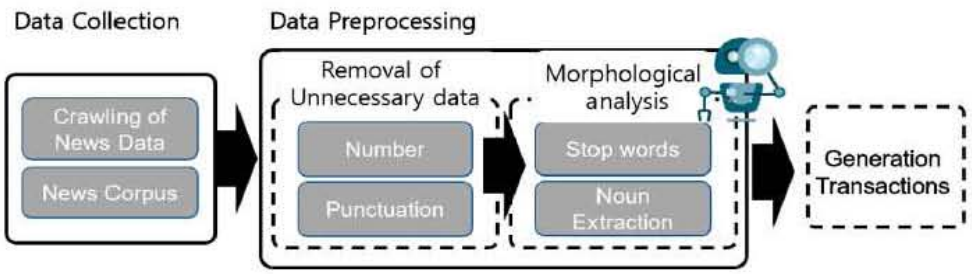
● TF-IDF 기반 순위 스코어를 이용한 연관 그래프의 최적화 프로세스



■ **교통 토픽 데이터 수집 및 전처리**

- 수집을 위한 웹 페이지는 국토교통부의 뉴스소식 보도자료[14]를 이용. 교통물류 토픽에는 사고, 자율주행, 교통 법규 등 다양한 교통 관련 정보들이 포함된 콘텐츠가 존재
- 기사 목록 페이지의 URL 주소에 대한 패턴을 분석하여 다수의 목록 페이지에 접근하여 연결되어 있는 모든 기사 본문의 URL 주소를 파악
- 뉴스의 본문이 들어있는 페이지 주소의 html 파일을 수집하여 html 파일에서 필요로 하는 기사의 본문을 가져오기 위해 별도의 파싱 과정이 필요함
- html의 기사 콘텐츠에 해당하는 태그의 클래스 명을 찾아 파싱하여 교통관련 뉴스 본문 데이터를 수집. 이는 해당 페이지의 bd_view_cont 클래스 명을 가진 div 태그 내의 본문 추출함
- 수집한 소셜 데이터에서 지식 그래프를 생성하기 위해 코퍼스로 변환하고, 뉴스 코퍼스는 해당 뉴스 생성 날짜, 카테고리, 뉴스 출간 회사, 주제와 원문, URL로 구성함
- 수집한 코퍼스는 분석의 정확도를 높이기 위해 전처리 과정을 진행함. 전처리는 뉴스 코퍼스에서 연관규칙을 적용하기 위해 소문자 변환, 구두점과 불용어 제거 등의 진행하여 문서 용어 행렬에서 트랜잭션을 구성

● 뉴스 코퍼스에서 전처리 과정



- 전처리는 불필요한 뉴스 생성 날짜, 카테고리, 뉴스 출간 회사와 URL 데이터를 삭제하며, 전처리 결과는 카테고리, 뉴스 주제와 원문의 내용으로 구성한 코퍼스이며, 뉴스 코퍼스에서 추출된 단어로 트랜잭션을 생성하고, 각 뉴스 주제와 원문에서 단어를 추출하여 형태소 분석을 진행함
- 형태소분석은 뉴스 코퍼스에서 문장부호, 숫자, 특수문자 등의 불용어를 제거하여 단어 추출. 불용어는 색인 단어로서 의미가 없는 단어이기 때문에 벡터로 변환된 매트릭스에서 제거하여, 추출된 뉴스 문서별 명

14) Ministry of Land, Infrastructure and Transport. https://www.molit.go.kr/.

사 리스트 단어들을 이용하여 단어들을 알맞은 형식의 트랜잭션으로 변환하는 작업을 수행하고 각 트랜잭션의 행마다 ID 값을 부여하여 각 트랜잭션을 라벨링 진행함

● 코퍼스에서 전처리 과정을 거친 트랜잭션

Transaction ID	Itemset
1	conflict, improve, initiate, discussion, view, open, common, relation, nation, attitude, year, content, history, Olympic, recognition, attention, ...
2	director, hold, game, stadium, gold medal, commemoration, north, south, attainment, battle, victory, national, team, Laos, start, football, India, ...
3	initiation, public, relations, related, content, effort, Tokyo, seeking, problem, missing, reflection, reparation, court, analysis, case, health, ...
4	director, hold, game, stadium, gold medal, commemoration, north, south, attainment, battle, victory, national, team, Laos, music, performance, ...
5	health, studying, performance, Netherlands, ability, broadcasting, growth, start, deterioration, Jansons, orchestra, music, bus, traffic, ...
6	possible, gas, emphasis, attack, science, global, finance, effort, news, achievement, response, city, car, expressway, licence, store, guest, ...
7	price, game, economy, customer, soaring, record, reporter, dollar, help, marketing, store, sales, USA, press, analysis, time, shopping, stress, ...
...	...

- Transection ID가 '1' 일 때, conflict, improve, initiate, discussion, view, open, common, relation, nation, attitude, year, content와 같은 item이 transaction 내에 포함되어 있다는 것을 의미하고, k항목을 포함하는 데이터 집합은 잠재적으로 2k-1개의 빈발 항목집합을 생성할 수 있고 기하급수적으로 탐색 영역이 커지며, 빈발 항목집합의 계산 복잡도를 줄이기 위해, 후보 항목집합의 개수를 줄이는 방법을 사용함

▪ 마이닝을 이용한 연관단어 지식 그래프 추출

- 연관단어 지식 추출을 위해 마이닝을 이용하여 연관규칙 발견하며, 연관규칙의 결과는 그래프로 표현하여 시각화를 진행하고, 동시에 예측할 반응변수를 어떠한 값으로 설정하여 단어간의 연관성을 파악할지에 대해 불확실한 측면이 존재하기 때문에 연관규칙 알고리즘을 사용하여 특별히 반응변수를 설정하지 않고 단어 간의 관계를 지지도와 신뢰도를 기준으로 파악할 수 있도록 함
- 생성된 트랜잭션을 바탕으로 연관규칙 알고리즘을 적용하여 최소지지도가 0.06, 신뢰도가 0.1 이상인 빈발 항목 헤더 테이블을 생성하고, 최소지지도와 신뢰도는 0~1의 값 사이에서 결정함

● 빈발항목 헤더 테이블

Item	Support	Count
news	0.5370112	769
local	0.3163408	453
time	0.2688547	385
relation	0.2632682	377
communication	0.2108939	302
police	0.2094972	300
report	0.2067039	296
government	0.2067039	296
reporter	0.1927374	276
news, local	0.1857542	266
America	0.1780726	255
middle	0.1745810	250
Trump	0.1717877	246
occur	0.1710894	245
Seoul	0.1682961	241
traffic	0.1658888	240
pot-hole	0.1658763	238
bridge	0.1648232	234
express	0.1623455	223
congestion	0.1593023	208
fog	0.1364892	186
...	...	...

- 최소지지도와 신뢰도의 설정 기준은 기존 연관규칙 데이터에서 불필요한 데이터셋을 제거하고, 연관단어를 충분히 생성할 수 있는 정도의 값으로 반복적인 실험을 통해 최소지지도와 신뢰도에 대한 최적의 값을 결정 단어의 연관 정도를 분석하여 연관성이 있는 단어 저장함
- 첫 번째 행의 'news' 는 지지도 0.5370112를 만족하며 문서 내에서 769번 등장했음을 의미함
- Apriori 알고리즘 기반 연관규칙은 규칙과 지지도, 신뢰도, Count로 구성하고, Count는 규칙에 사용된 단어들의 횟수를 나타내며, 지지도는 Lhs와 Rhs가 동시에 나오는 확률을 뜻하므로, 같은 지지도와 Count값이 나타나지만, 신뢰도는 연관규칙을 생성하는 단어가 Lhs와 Rhs의 순서가 교환되면 규칙생성에 사용되는 단어가 같기 때문에 지지도와 Count 값은 같지만 신뢰도 값은 다르게 나타남

● Apriori 알고리즘 기반 연관규칙 알고리즘 결과

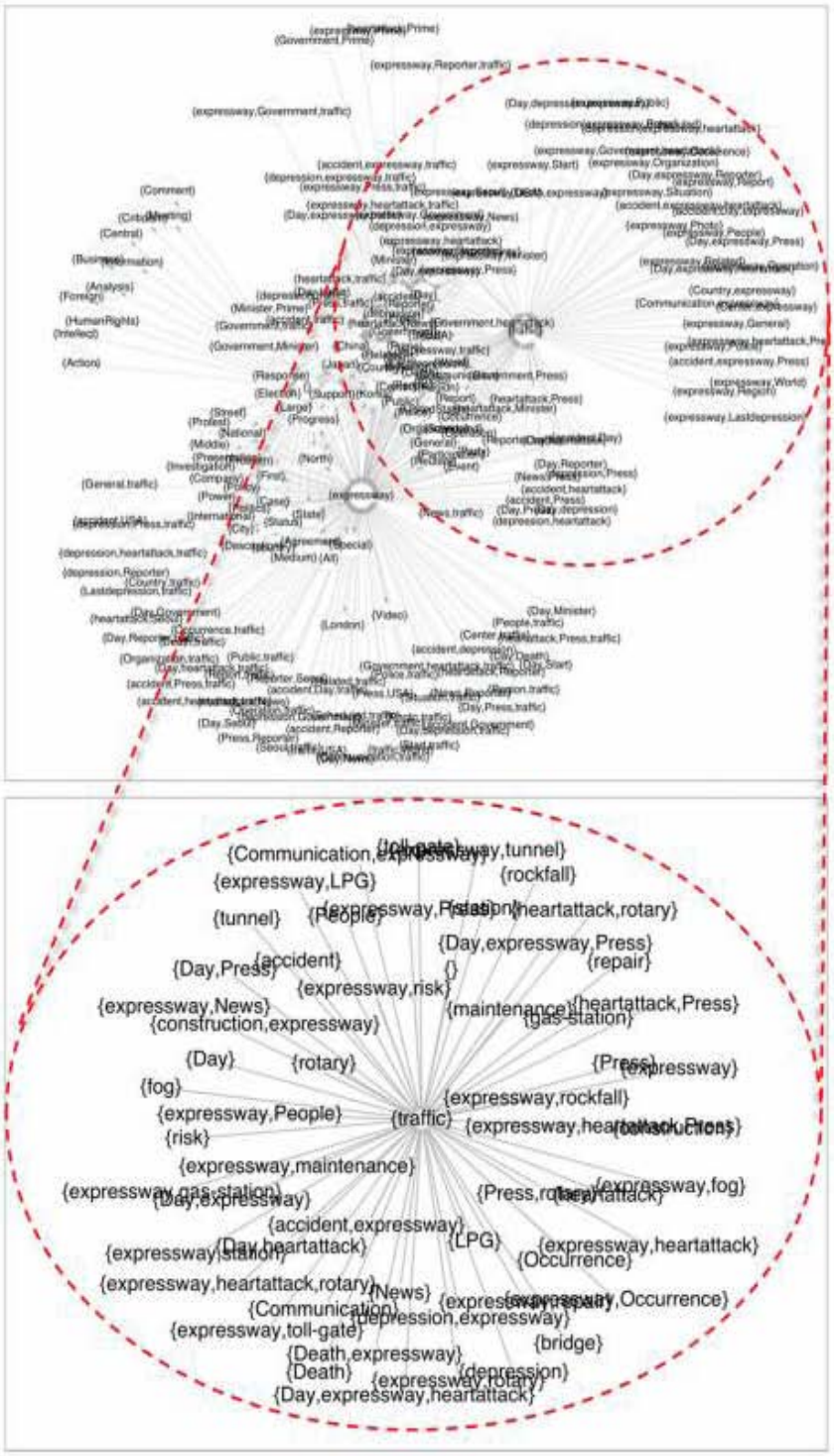
Lhs => Rhs	Support	Confidence	Count
local => news	0.18575	0.58719	266
news => local	0.18575	0.34590	266
time => local	0.16829	0.62597	241
local => time	0.16829	0.53200	241
time => news	0.16689	0.62077	239
news => time	0.16689	0.31079	239
relation => news	0.14315	0.54376	205
news => relation	0.14315	0.26658	205
seoul => news	0.12220	0.72614	175
news => seoul	0.12220	0.22756	175
police => news	0.11592	0.55333	166
news => police	0.11592	0.21586	166
communication => news	0.11522	0.54635	165
news => communication	0.11522	0.21456	165
reporter => news	0.11451	0.58320	164
news => reporter	0.11451	0.49093	164
traffic => speed	0.11449	0.61239	160
speed => traffic	0.11449	0.41233	160
distance => time	0.11389	0.59234	159
time => distance	0.11389	0.31456	159
gas-station => explosion	0.10998	0.56133	155
explosion => gas-station	0.10998	0.32331	155
expressway => speed	0.10967	0.65333	152
speed => expressway	0.10967	0.45467	152
...	...	...	...

- 첫 번째는 Lhs가 local이고, Rhs가 news로 규칙이 생성되었을 때 지지도는 0.18575, 신뢰도는 0.58719, 횟수는 266이 나타나며, 두 번째는 Lhs가 news이고, Rhs가 local로 규칙이 생성되었을 때 지지도는 0.18575, 신뢰도는 0.34590, 횟수는 266이 나타남. 첫 번째 행과 두 번째 행의 지지도와 횟수는 같게 나타나지만 Lhs와 Rhs가 변경되었기 때문에 신뢰도는 다르게 나타남
- 연관규칙을 기반으로 단어 X와 Y의 연관성을 나타내는 X => Y 표시를 기반으로 edge와 vertex를 생성하며, X => Y의 의미는 단어 X가 등장할 때 단어 Y가 나타날 수 있는지에 대한 연관성을 나타내며 그래프를 구성하면 다음과 같은 뉴스기반 단어 연관 그래프가 생성됨[15)]

- 연관규칙 알고리즘을 이용한 연관단어 지식 그래프는 방향 그래프를 통해 지식 그래프를 추출하고, 노드는 각 단어들을 나타내며, 엣지는 그래프의 방향성을 나타내며, 방향으로 각 단어 간의 연관성을 나타냄
- 그래프에서 생성된 규칙의 수는 약 2,400개이며, 생성된 규칙 중에서 양방향 화살표는 단어 간의 상호간 연관규칙이 존재함을 의미하기 때문에 규칙의 수는 2,400개이지만 생성된 지식의 수는 더 방대할 수 있음. 특정 단어가 여러 단어와 연관관계를 맺고 있기 때문에 단어의 중복으로 인한 규칙과 불필요한 지식이 포함되고, 시각화에 대한 효율성이 낮아지기 때문에 지식 획득에 한계가 존재함

● 연관규칙 알고리즘을 이용한 연관단어 지식 그래프

15) 권혜정, 김민정, 류성은, 정경용, 클러스터링 기반 중심점 분석을 이용한 실시간 교통안전등급 산출모델, 2021 한국인터넷정보학회 춘계학술발표대회, pp. 73-74, 2021.



■ **TF-IDF를 이용한 연관지식 그래프 최적화[16]**

- 생성된 지식 그래프에서 불필요하고 중요도가 낮은 단어가 다수 포함되어 있으며, Apriori 알고리즘의 특성상 하나의 단어에 대한 Term Frequency가 고려되지 않고 하나의 문서에 해당 단어가 여러번 등장하여도 1번으로 계산한다는 문제점을 해결하기 위해 연관단어 그래프를 최적화하여 더 정제된 모델을 높은 속도로 생성하기 위해 추출한 단어에 대한 TF-IDF 값을 계산

- TF-IDF값을 적용하기 위해 1,700개의 교통안전 뉴스 데이터의 단어에 대한 문서 용어 행렬(document term matrix)[17]을 생성하여 TF×IDF를 계산한 Matrix를 생성
- 문서 용어 행렬(document term matrix)에서 단어 사이의 상호작용을 표현하지 못하는 문제점을 해결하기

16) 류성은, 김현진, 구병국, 권혜정, 박찬홍, 정경용, 도로 위험 탐지를 위한 데이터 편향성 최적화 기반 연관 추론 모델, 한국융합학회논문지, 11(9), 1-6, 2020.

17) 권혜정, 김민정, 강지수, 백지원, 정경용, 교통안전을 위한 구조적 시각 유사도를 이용한 비지도 이상치 탐지, 2020 한국인터넷정보학회 추계학술발표대회, pp. 81-82, 2020.

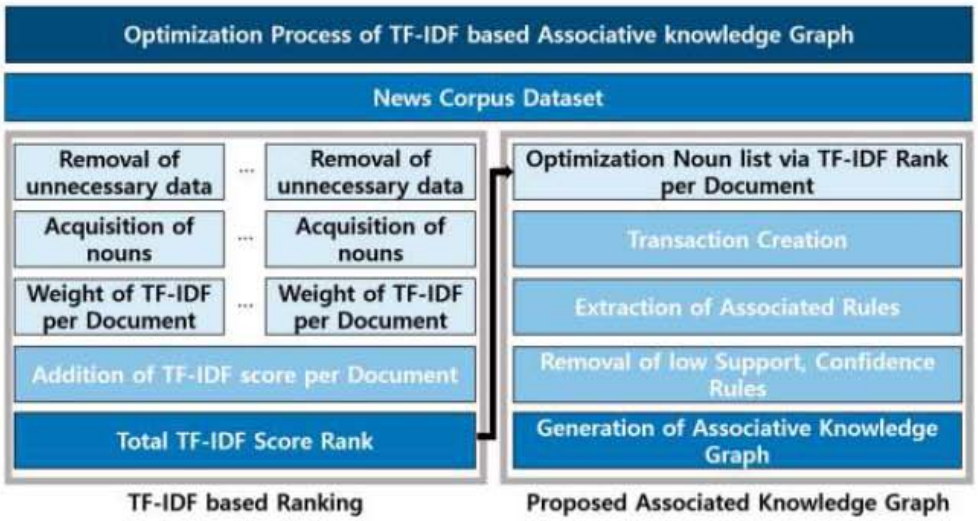
위해 TF-IDF가 계산된 Matrix에서 단어당 모든 문서의 TF-IDF를 합하여 TF-IDF를 이용한 순위가 부여

● TF-IDF 순위 스코어의 결과

TF-IDF Rank	Word	TF-IDF Rank Score
1	traffic	35.206371
2	depression	31.395520
3	video	21.520154
4	police	21.021509
5	stress	19.089421
6	accident	18.581047
7	injury	18.428506
8	organization	18.072522
9	bus	18.026231
10	photo	17.912217
11	expressway	17.878551
12	president	16.859654
13	psychotherapy	16.203685
...	...	...

- 생성된 TF-IDF 스코어 순위를 기반으로 하여 상위 20%의 단어를 바탕으로 연관 그래프와 비교하며, 뉴스 데이터에서 추출한 단어 중 상위 20%의 단어가 존재하지 않는다면 해당 단어를 제거하는 과정 진행함
- 전처리 과정의 결과로 생성된 트랜잭션을 바탕으로 연관단어 규칙을 생성하여 최종적으로 주요 단어로 생성된 연관단어 지식 그래프를 생성함

● TF-IDF를 이용한 연관지식 그래프의 최적화 과정



- 첫 번째 단계로 TF-IDF Rank Table을 생성하기 위해 데이터를 수집 및 전처리 진행하고, TF-IDF 가중치를 계산하기 위해 뉴스 코퍼스 데이터에서 단어 추출함. 각각의 뉴스 문서에서 TF-IDF 가중치가 부여된 단어들을 추출하여 모든 뉴스 문서에 적용하여 각 단어의 가중치를 합산하여, 각 단어의 가중치 합산 결과는 TF-IDF Rank Table이 생성됨

- 두 번째 단계로 연관규칙을 생성하기 위해 데이터 전처리 과정 진행 후 뉴스 코퍼스에서 URL과 같은 불필요한 데이터열을 삭제하고, 형태소분석을 하여 단어 데이터를 추출하고, 단어 데이터가 TF-IDF Rank Table의 하위권에 존재할 경우 해당 단어는 제거함
- 최종적으로 TF-IDF Rank 행렬 기반 최적화된 트랜잭션을 생성하게 됨

● 최적화된 트랜잭션 생성 알고리즘

Algorithm 1 Optimized Transaction Generation Algorithm

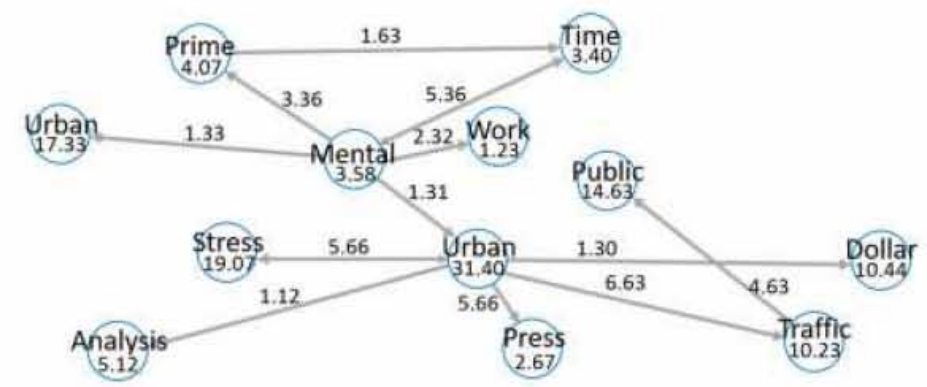
```
Input : News Stream Data → NewsData[k]
Output : Optimized Transaction → OptimizedTransaction[l]
Step 1: TF-IDF Ranking Table Generation
RankTable ← NULL // TF-IDF RankTable
for i is number of NewsData[k] do
    TF_Nouns ← extractNoun(NewsData[i])
    TF_Corpus ← Corpus(TF_Nouns) // make TF_Nouns list to Corpus data
     removeUnnecessaryData(TF_Corpus)
    // Create DocumentTermMatrix from TF_Corpus to calculate TF-IDF Weight
    TF_DTM ← DocumentTermMatrix(TF_Corpus)
endfor
TF_DTM ← WeightTFIDF(TF_DTM)
// Add TF-IDF Weights per Column to sum up TF-IDF of single News Data
for i is number of column in TF_DTM do
     RankTable[i] ← Sum(TF_DTM[i])
endfor
// Sort RankTable by Score
RankTable ← Sort(RankTable, RankTable.TotalColumnScore)
Step 2: TF-IDF based Optimized Transaction Creation
for i is number of NewsData[k] do
    Nouns[i] ← extractNoun(NewsData[i]) // extract words from a row of news data
    for j is number of words extracted in Nouns[i] do
        for k is number of words in top 20% of RankTable do
            // remove word if it cannot be found in the top 20% RankTable
            if (Nouns[j] != RankTable[k])
                Remove(Nouns[j])
        endfor
    endfor
    NewsCorpus ← Corpus(Nouns[i])
    removeUnnecessaryData(NewsCorpus)
    OptimizedTransaction[i] ← Transaction(NewsCorpus)
endfor

return OptimizedTransaction
```

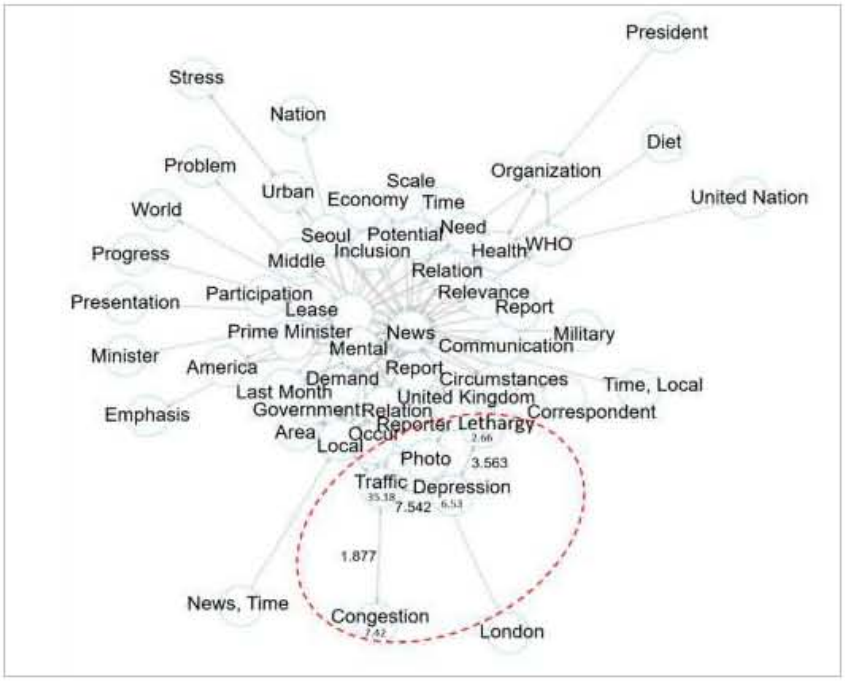
- 생성된 트랜잭션을 기반으로 연관규칙 알고리즘에 적용하여, 단순한 연관규칙에서 잠재적인 연관관계 및 지식을 추출하고 데이터에 담긴 의미를 효과적으로 추출하고 관찰하기 위한 모델이 필요하기 때문에 정제된 연관규칙을 추출하기 위해 Pruning을 진행하고, 생성한 연관규칙을 기반으로 데이터를 시각화함

- 그래프는 개체 사이의 관계를 나타낼 수 있으며 종류는 단방향 그래프와 양방향 그래프로 구분함
- 그래프의 edge 가중치는 연관규칙의 Lift 값을 이용하고, 중요 단어를 통한 연관규칙 생성을 위해 vertex는 단어의 중요도를 나타내는 TF-IDF 가중치를 이용함

● 연관규칙 기반의 데이터 시각화 결과



● TF-IDF를 이용한 최적화된 연관지식 그래프 결과



- 교통사고 관련 뉴스 코퍼스에서 발생하는 건강 문제에 대한 정보를 추출하기 위해 단어 토픽을 모델링하여, 교통안전 및 건강 토픽의 상호관계에서 이머징 헬스 리스크 예측에 대한 지식베이스로 확장 가능함

■ 감성 유사도 가중치를 이용한 이머징 리스크 탐지를 위한 워드 임베딩 기반 교통 문서 분류 모델[18]

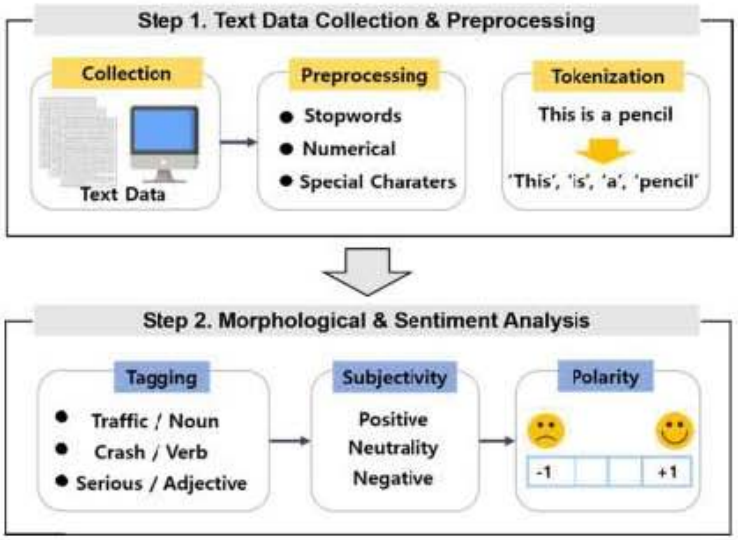
- 감성분석(Sentiment Analysis)은 텍스트가 내포하고 있는 긍정과 부정의 정도에 대한 극성값을 파악하여 텍스트의 감정 상태를 파악함. 감정 상태 분석을 위해서 감성 어휘들의 극성값으로 이루어진 감성사전 구축하며, 텍스트로 이루어진 데이터들을 전처리한 뒤 해당 단어가 가지고 있는 감성을 정량화하여 표현함
- 감성분석은 감성사전 기반 분석 방법과 기계학습 기반 분석 방법이 존재하며, 감성사전 기반 감성분석은 긍/부정값을 수치화하여 나타냄으로써 단어의 감성을 분석할 수 있도록 하지만 감성사전에 측정해야하는 단어가 존재하지 않을 경우 감성을 측정하기 어렵고, 단어의 유무와 단어의 극성 정도에 따라 분석 결과가 달라지기 때문에 신중한 감성사전의 구축이 요구됨

- 감성사전 구축을 위해 머신러닝과 딥러닝을 이용한 지도학습 방법이 감성분석에 자주 사용되고 있음. 지도학습은 사용자가 부여한 라벨을 통해 학습하며 이를 통해 답을 유추해 나가는 과정을 의미하며, 분서의 긍부정도를 Label로 부여함으로써 분류기는 스스로 학습하고, 결과를 판단함

18) M. J. Kim, J. S. Kang, K. Chung (2020) Word-Embedding-Based Traffic Document Classification Model for Detecting Emerging Risks Using Sentiment Similarity Weight, IEEE Access, 8, 183983-183994.

- 감성 어휘사전 기반 모듈로 TextBlob, VADER(Valence Aware Dictionary and sEntiment Reasoner), SentiWordNet 등이 사용하여 진행함
- 감성사전 구축은 텍스트 데이터 수집 단계와 형태소 및 감성분석 단계로 구성됨. 첫 번째 단계는 텍스트 데이터 수집 및 전처리 단계이며, 전처리 과정은 데이터들을 수집하고 문서화하며, 불필요한 단어들을 제거 또는 토큰화 진행함. 두 번째 단계는 수집한 텍스트 데이터로부터 형태소분석과 감성분석을 수행하며, 문서화된 데이터를 형태소를 분석하여 태깅한 뒤 감성 분석하고, 또한 단어의 감성 정도를 파악하여 긍정, 중립, 부정으로 분류하고, 정량화함
- 기존의 감성분석 방법들은 하나의 감성사전을 통해 어휘들의 극성값을 구성하며, 하나의 감성사전에 존재하는 단어들에 국한해서 긍부정도를 분류하기 때문에 감성사전에 특정 단어가 존재하지 않을 경우 긍부정도를 분류하지 못하는 한계가 존재하기 때문에 기존의 감성사전을 확장하여 감성사전에 단어가 존재하지 않는 경우에도 긍정/부정을 산출할 수 있는 방법이 필요함

• 감정사전 구축 프로세스

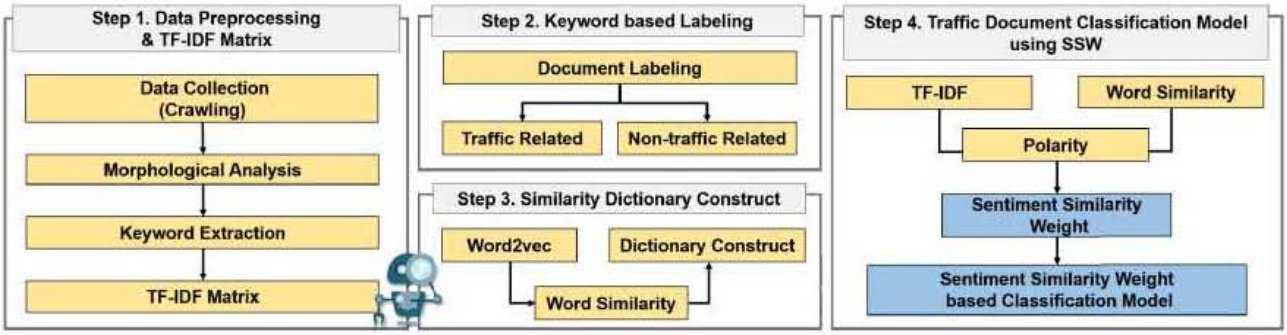


- 워드 임베딩은 문장에 등장하는 단어들의 맥락을 파악하여 벡터의 값으로 변환하며, 워드 임베딩 방법으로는 GloVe, Fasttext, Word2vec 등이 존재함. word2vec은 사용자가 지정한 window 안에서만 학습이 진행되기 때문에 문장 전체의 동시출현빈도를 고려하지 못하는 단점이 있으나 GloVe(Global Vectors)는 전체 단어들의 동시출현 확률을 이용하여 의미 있는 단어의 구조를 형성하지만 단어가 많아질수록 동시출현단어(Co-occurrence word) 매트릭스의 크기와 계산량이 증가하는 단점이 있음
- Fasttext는 페이스북이 제안한 워드임베딩 방법이며, 하나의 단어에도 여러 단어들이 존재하는 것으로 가정하여 단어의 벡터값을 n-gram의 합으로 구성함으로써 사전에 존재하지 않는 단어의 벡터를 생성하고, 학습시간이 빠름
- Word2vec는 워드 임베딩 방법 중 하나로 구글에서 기존의 NNLM(Neural Network Language Model)의 성능을 개선하여 제안한 모델이며, 비슷한 위치에 존재하는 단어들은 유사한 의미를 지닌다는 분포 가설에 근거하여 단어 간의 의미를 유추하고, 단어를 고차원의 좌표에 표현함으로써 단어에 벡터를 부여함으로써, 단어 사이의 연산이 가능하기 때문에 단어 간의 유사도 계산이 가능함
- Word2vec은 CBOW(Continuous Bag Of Word)방법과 Skip-gram 방법으로 주로 사용됨. CBOW는 주변단어들을 input으로 사용하여 주어지지 않은 중심단어를 예측하는 방법이며, 학습 시간이 빠르다는 장점이 있고, Skip-gram은 중심단어를 input으로 하여 주변 단어들을 예측하는 방법이며, 맥락의 수만큼 손실을 계산해야 하기 때문에 학습 시간이 느림

- CBOW인 경우 중심단어의 벡터값을 업데이트하기 때문에 학습의 효율성은 CBOW 방식보다 우수하게 평가되며, Skip-gram은 윈도우의 크기의 2배만큼 업데이트하기 때문에 같은 윈도우 크기를 갖더라도 학습량이 수 배 이상 차이가 존재함으로 일반적으로 학습의 효율성이 높은 대부분 Skip-gram 방식을 사용함

- 텍스트 데이터에서는 정보 획득이 가능하지만 텍스트를 통해 위험 정도를 인지하고 예측하기에는 어려움이 존재하고, 다양한 텍스트 문서로부터 교통과 관련된 내용을 포함한 교통 문서를 수집하고 찾아낸 교통 문서를 통해 이머징 리스크를 탐지하기 위한 방법이 필요함
- 감성 유사도 가중치를 구성하여 감성 사전에 없는 단어라도 가중치를 부여하여 분류에 유의미하게 사용되도록 함

• 감성 유사도 가중치를 이용한 이머징 리스크 탐지를 위한 워드 임베딩 기반 교통문서 분류 프로세스



- 크롤링을 통해 비정형 데이터를 수집하고 전처리를 진행한 뒤, TF-IDF 가중치를 통해 매트릭스를 구성함. 수집된 문서에서 불필요한 단어를 제거하고 형태소분석 과정을 진행하여 TF-IDF 가중치를 가지는 중요 단어들을 추출함

- 문서에서 추출된 중요 단어들을 통해 문서를 라벨링하여 교통 관련 데이터와 non-traffic 데이터를 분류함. Word2vec를 통해 단어를 벡터화하고 유사도를 계산하여 사전을 생성함
- 키워드의 TF-IDF와 Similarity, Polarity을 통해 Sentiment Similarity Weight를 도출하여, 감성사전에 포함되지 않은 단어들은 가장 높은 유사도를 지니고 있는 단어들로 대치함으로써, Sentiment Similarity Weight를 이용한 분류 모델을 시스템에 적용함

■ **TF-IDF를 이용한 데이터 수집 및 키워드 추출**

- 교통문서에서 이머징 리스크를 탐지하기 위해 뉴스속보 데이터에 대한 크롤링을 진행함. TBN 한국교통방송[19]에서 제공하는 데이터를 정치, 경제, 사회, IT과학, 교통 등 토픽별로 수집하여 문서화함으로써 텍스트 데이터 수집함. 텍스트 데이터에서 문서가 작성된 날짜와 시간, 카테고리, 제목, 본문, 링크 등의 분석에 필요한 데이터만 수집함
- 크롤링한 뉴스 텍스트 데이터에는 카테고리가 교통이 아니지만 교통과 관련된 문서들이 존재하기 때문에 수집한 문서의 카테고리를 다시 판단할 필요성이 있음. 문서가 교통과 관련되었는지 판단하기 위해 문서의 키워드를 추출하고, 추출된 키워드를 기반으로 문서가 교통 관련 문서인지에 대한 binary 라벨링을 진행함

- 크롤링한 문서 안에 존재하는 대표 키워드를 추출하기 위해 형태소 분석과정을 진행하여 각 문서마다 토큰화를 수행한 뒤 무의미한 단어라고 판단되는 접속사, 조사, 숫자, 특수 부호 등을 포함하는 불용어 처리함. 전처리한 데이터에서 단어의 중요도에 따라 주요 키워드를 추출하기 위해 TF-IDF(Term Frequency - Inverse Document Frequency) 가중치 매트릭스를 구성함. TF-IDF는 단어 빈도에 역문서 빈도를 곱하여 단순하게 다빈도로 나타나는 키워드의 가중치를 조정하여 의미있는 키워드만 높은 중요도를 보이도록 함

- 키워드 추출에 많이 사용되는 형태소는 명사이지만, 감성어휘는 다른 품사에도 많이 분포되어 있기 때문에, 한가지의 품사만을 이용하여 감성분석을 수행하기 어렵기 때문에 정확하고 효과적인 감성분석을 위해 감성어휘에 많이 포함되어 있는 동사, 형용사도 추출하여 사용하고, 추출한 키워드를 리스트화하여 단어 품사에 따른 명사, 동사, 형용사에 대한 각각의 TF-IDF 가중치 매트릭스 구축함

19) Traffic Broadcasting Network, https://www.tbn.or.kr/.

• 문서에서 명사 단어의 TF-IDF 가중치 매트릭스

NO.	sa-go (accident)	gyung-chal (police)	sool-chi-han (drunk)	wi-hum (risk)	cha-ryang (vehicle)	…
Doc_{54}	0	0.315	0	0.210	0	…
Doc_{72}	0.446	0.147	0.179	0.238	0	…
Doc_{103}	0	0.526	0	0	0.364	…
Doc_{128}	0.741	0	0	0.512	0.770	…
Doc_{139}	0.678	0	0.199	0	0.541	…
Doc_{165}	0.567	0.631	0	0.279	0.128	…
…	…	…	…	…	…	…

- 첫 번째 행은 한국어 단어들의 발음을 나타내고, 괄호 안은 단어들의 영어 번역을 나타냄.
- Doc54에서 "sago" (accident)의 TF-IDF 가중치는 0이고, Doc128에서는 0.741이며, accident는 Doc54에서 의미 없는 단어이지만 각 문서마다 단어들의 중요도가 다르기 때문에 Doc128에서는 유의미한 단어로 나타남
- TF-IDF 가중치 매트릭스를 구축함으로써 문서에 존재하는 주요키워드를 파악하고 해당 문서에서 특정 키워드의 중요도를 산출함
- 문서의 주요 키워드 추출을 통해 문서의 토픽을 파악한 뒤 교통과 관련이 있는 문서들은 1로 표기하고, 교통과 관련이 없는 문서들은 0으로 하여 이진 라벨링을 진행하기 위해 국토교통부에서 제공하는 국토교통용어사전[20]의 교통 키워드를 기반으로 관련문서를 라벨링함
- 라벨 1은 'open highway', 'traffic safety law revision', 'traffic accident', 'road congestion' 등과 같은 교통과 관련있는 문서이며, 라벨 0은 교통과 전혀 무관한 문서들로 나타냄
- 교통과 관련된 문서로 판단되어 Label 1로 라벨링된 문서는 location을 수집하고, 교통과 관련이 없는 라벨 0인 문서들은 수집하지 않고 삭제하고, 키워드 기반의 라벨링을 수행하기 때문에 문서의 카테고리 기반이 아닌, 문서의 실질적인 내용에 기반하여 보다 정확한 문서의 분류 가능

• 문서의 추출된 키워드와 교통 라벨링 결과

NO.	Keyword	Time	Location	Category	Label
1	Gangdong-gu/NNP, eo-rin-e(children)/NNG, bo-ho(protection)/NNG, gwa-sok(overspeeding)/NNG, alarm/NNG, sin-ho(signal)/NNG, system/NNG, gang-hwa(enhance)/NNG, ppa-reu-da(fast)/VV...	2020-01-15 21:56	Gangdong-gu	Society	1
2	Incheon /NNP, ji-yeok(area)/NNP, bal-sang(occurrence)/NNG, o-to-ba-e (motorcycle) /NNG, sa-go(accident)/NNG, jeung-ga(increase)/NNG, na-ta-na-da(appear)/VV, gyo-tong(traffic)/NNG, (je-han)limit/NNG...	2020-01-31 13:26	Incheon	Traffic	1
3	a-chim(morning)/NNG, do-ro(road)/NNG, ja-dong-cha(car)/NNG, taxi/NNG, chung-dol(collision)/NNG, gyo-tong(accident)/NNG, un-haend(driving)/NNG, unjeongja(driver)/NNG, juk-da(die)/VV, byeong-won(hospital)/NNG, tap-seung-ja(passenger)/NNG, sim-gak(serious)/NNG, son-sang(damage)/NNG, e-dong(transfer)/NNG...	2020-02-19 11:27	Seobuk-gu, Cheonan-si, Chungcheongnam-do	Traffic	1
4	audio/NNP, sa-yong-ja(user)/NNP, jeung-ga(increase)/NNG, hae-woi(overseas)/NNG, audio-chaek(audiobook)/NNP, hoi-sa(company)/NNG, ttwi-eo-deul-da(dive)/VV, service/NNG, so-ri(sound)/NNG, da-yang-seong(diversity)/NNG...	2020-01-01 10:31	-	IT Science	0
5	do-ro(road)/NNP, gyo-tong(traffic)/NNG, kwon-han(authority)/NNG, dan-sok(regulation)/NNG, camera/NNG, seol-chi(installation)/NNP, jae-tam-saek(research)/NNG, go-sok-do-ro(highway)/NNG…	2020-02-07 9:40	Gyeongbu Express, Seoul toll booth	Traffic	1
6	truck/NNG, un-jeon(driving)/NNG, gyo-tong(traffic)/NNG, bul-bit(light)/NNG, yeon-swae(concatenation)/NNG, chung-dol(collision)/NNG, (Jung-bu nae-ryuk go-sok-do-ro)Central Inland Expressway /NNP, sa-go(accident)/NNG...	2019-12-24 11:34	Central Inland Expressway	Traffic	1
7	fastfood/NNG, coffe/NNP, sang-jum(store)/NNP, ja-dong-cha(car)/NNG, ta-da(ride)/VV, chaek(book)/NNG, bil-li-da(borrow)/VV, ban-nap(return)/NNG, convenient/NNG, sa-yong-ja(user)/NNP…	2020-01-01 15:14	-	Society	0
8	ji-ha-chul(subway)/NNG, bal-saeng(operation)/NNG, jung-dan(interruption)/NNG, boramae/NNP, yeok(station)/NNG, ddu-nam(departure)/NNG, won-in(cause)/NNG, mum-choo-da(stop)/VV...	2020-01-22 23:42	Boramae Station	Traffic	1
…	…	…	…	…	…

20) Ministry of Land, Infrastructure and Transport of Korea, http://www.molit.go.kr/.

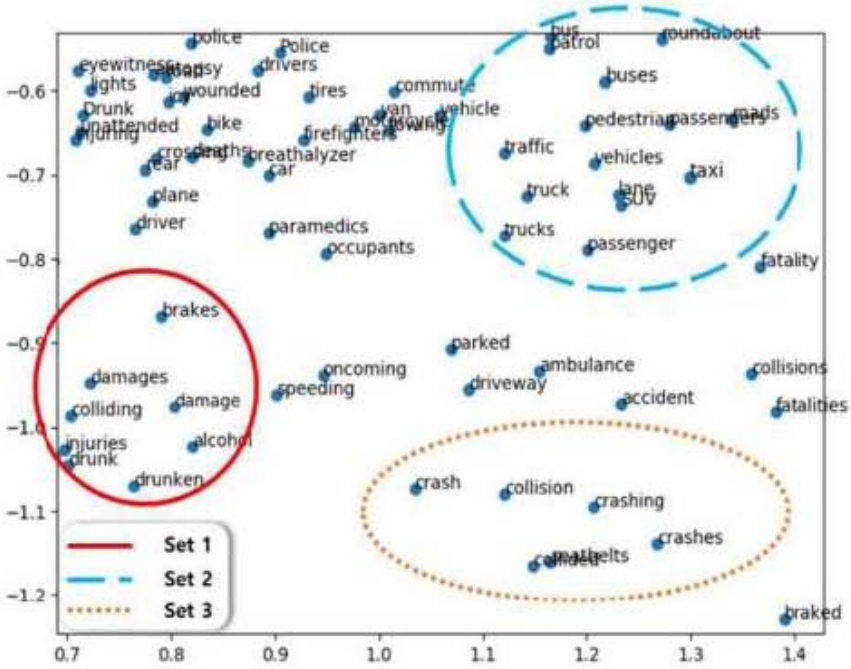
- 크롤링된 문서는 전처리되어 Contents, Time, Location, Category, Label의 반정형데이터로 구성됨. Contents는 문서에서 사용된 단어와 형태를 나타냄. NNG는 일반명사, NNP는 고유명사, VV는 동사를 의미함. Time은 문서가 수집된 날짜와 시간을 의미함. Location은 문서에서 나타난 교통 이벤트의 발생 위치와 범위를 나타냄. Category는 정치, 사회, 경제, IT 과학, 교통 등으로 범주화하여 나타냄
- 기존의 문서에서 교통관련 내용이 포함되더라도 카테고리가 교통이 아니라면, 교통문서로 분류되지 않는 단점이 있지만 문서의 키워드를 고려한 방법을 통해 문서의 클래스가 교통이 아니라도 교통관련 내용의 유무에 따라 교통 문서로 분류가 가능하며, 교통 카테고리가 아닌 다른 카테고리에서도 교통관련 문서 수집이 가능하기 때문에 한정된 카테고리에서 문서를 수집하는 한계점을 보완하고 정확성을 높이는 것이 가능

■ **Word2vec를 사용한 코사인 유사도 기반 키워드 유사도**

- 단어에는 형태가 서로 다르지만 뜻이 같은 다수의 이음동의어가 존재하며, 이음동의어를 통해 사용되지 못하는 단어가 존재할 때, 단어 간의 의미적 유사성을 발견하여 대치가 가능함
- 단어 간의 유사도를 계산하기 위해 word2vec 기반으로 추출된 키워드를 벡터화하고, 유사도 결과를 통해 사전을 구축하기 위해 크롤링한 문서 데이터를 사용하며. 키워드를 벡터화하기 위한 Word2vec는 문장 내의 맥락을 통해 단어를 예측함

- 유사도 사전 구축 데이터의 불용어 처리는 하지 않으며, 학습되는 단어의 다양성을 위해 교통 문서만을 사용하지 않고 경제, 정치, 사회 등의 여러 분야와 관련된 문서들을 사용함

• word2vec 기반의 워드 임베딩 결과



- Set1, Set2, Set3은 의미적으로 유사한 키워드들이 군집되며, 각각 damage, crash, vehicle의 집합을 의미하며, 군집된 키워드는 유사도가 높을수록 근접한 곳에 위치함
- 단어 간의 유사도를 발견하기 위해 벡터 공간에서 코사인 유사도를 이용하며, 코사인 유사도는 벡터 사이의 코사인 각도를 통해 유사도를 측정하기 때문에 다차원의 공간에서 벡터들의 거리를 구하는 것이 가능함
- 코사인 유사도는 -1부터 1까지의 실수값을 가지며, 단어 간의 코사인 유사도가 -1에 근접할수록 의미가 반대됨을 의미하며, 1에 근접할수록 의미가 유사함을 의미함

- 코사인 유사도에서 W_k와 X_k는 단어의 벡터값을 의미하며, $S(W_k, X_k)$는 W_k와 X_k간의 코사인 유사도를 의미함
- Word2vec을 이용한 단어 간의 코사인 유사도 결과에서 Tn은 유사도 순서를 의미하며, 이는 각 단어 간의 유사도가 높은 순으로 정렬함

• 코사인 유사도의 식

$$S(W_k, X_k) = \frac{\sum_{k=1}^{n} W_k X_k}{\sqrt{\sum_{k=1}^{n} (W_k)^2} \times \sqrt{\sum_{k=1}^{n} (X_k)^2}}$$

• Word2vec을 이용한 단어 간의 코사인 유사도 결과

Tn	un-jeon(driving)		ja-dong-cha(car)		go-sok-do-ro(highway)		…
T1	un-haeng(race)	0.876	truck	0.913	do-ro(road)	0.884	...
T2	joo-haeng(run)	0.749	ja-dong-cha(automobile)	0.854	cha-do(roadway)	0.809	…
T3	un-jeon-ja(driver)	0.744	SUV	0.733	un-jeon-ha-da(drive)	0.751	…
T4	do-ro(road)	0.688	un-jeon-ha-da(drive)	0.724	hyu-sik(resting)	0.709	…
T5	gyo-tong(traffic)	0.632	ba-kwi(wheel)	0.768	neu-ri-da(slow)	0.692	…
…	…	…	…	…	…	…	…

- “un-jeon” (driving)와 코사인 유사도가 가장 높은 단어는 0.876의 유사도 값을 가진 “un-haeng” (race)을 나타내며, 두 번째는 0.749의 유사도 값을 가진 “joo-haeng” (run)을 나타내어, 유사도 사전을 구축함으로써 문서에서 존재하지 않는 주요키워드와 유사한 단어들을 도출함
- 단어 대치(imputation)를 수행하며, 단어 대치(imputation)는 문서에 존재하지 않는 경우 단어를 유사도를 기반으로 대체하는 방법으로, 코사인 유사도 결과를 통해 도출된 가장 높은 유사도를 나타내는 값 T1이 굉장히 낮은 값이 도출되는 worst case가 발생할 수 있음. T1이라고 하더라도 원래의 키워드와 T1이 의미적으로 유사하지 않은 경우가 존재함

- 일정 값 이상의 유사도를 가지는 단어들만 대치를 수행. 대치 임계값을 설정하기 위해 35,729개의 각 단어와 T1과의 유사성을 비교하기 위해 유사도를 0.1씩 줄여나가면서 단어와 T1이 유의미한 상관관계를 갖는지를 판단함

• 대치 임계값에 따른 단어와 T1이 의미적으로 일치할 확률

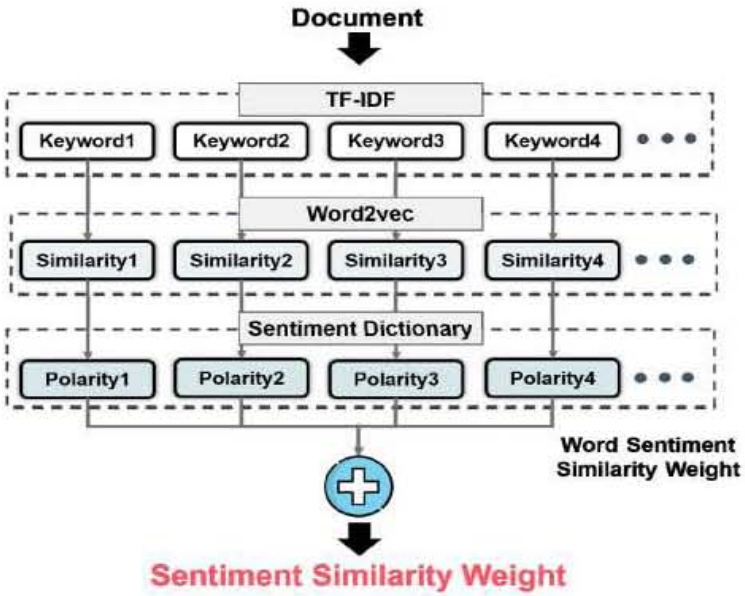
Threshold	1.0	0.9	0.8	0.7	0.6	0.5	0.4
Value	1.0	0.875	0.812	0.794	0.628	0.211	0.193

- Threshold는 단어와 T1간의 유사도 임계값을 의미함. Value는 각 임계값에서 단어와 T1이 의미적으로 유의미할 확률을 의미하고, Threshold가 0.5인 시점에서 cut off 되기 때문에 대부분 단어 간의 유사도가 0.5 이하일 때, 의미적으로 일치하지 않는다고 판단하여 대치 임계값을 0.5로 설정함. 단어의 유사도 기반 대치를 통해 유의미한 관계를 가진 단어들을 도출하는 것이 가능함

■ **이머징 리스크 탐지를 위한 감정 유사도 가중치**

- 이머징 리스크 탐지를 위한 SSW(Sentiment Similarity Weight)를 구축하기 위해 KOSAC(Korean Sentiment Analysis Corpus) 사용함. KOSAC은 16000여 개의 n-gram 형태의 형태소를 추출하여 데이터베이스에 저장하고, 긍정, 중립, 부정에 대한 -1에서 +1 사이의 값의 범위를 가지는 극성값을 구성하여 가중치 값을 산출
- 문서의 주요키워드에 따른 TF-IDF, Similarity, Polarity를 이용하여 단어들의 가중치 값을 생성하고, 문서에 존재하는 단어들의 합을 통해 최종적인 SSW를 도출함

• SSW의 도출 프로세스



- Sentiment Dictionary의 Polarity를 이용할 때 단어들이 감성사전에 존재하지 않으면 해당 단어를 사용할 수 없는 근본적인 문제점이 존재하기 때문에 감성사전에서 단어의 존재 여부에 따라 유사도와 극성값을 고려한 단어대치 방법을 사용함
- 대치 과정에서 단순히 가장 유사도가 높은 단어의 polarity를 바로 채택하면 단어간의 의미적 요소와 유사성을 고려하지 않아 가중치에 영향을 주기 때문에 단어간의 유사한 정도만큼만 값을 대치하기 위해 적용
- 단어 간의 유사도가 높을 경우, 의미적으로 유사하다고 판단되기 때문에 대치되는 단어의 극성 값이 고려되며, 유사도가 낮은 경우에는 대치되는 단어의 극성값이 작게 고려할 수 있기 때문에 단어의 유사도를 고려한 단어의 대치가 가능함

- 문서가 포함하고 있는 명사, 형용사, 동사 각각의 단어에 대한 감성 유사도 가중치 산출하여, 단어 w에 대한 WSSW(Word Sentiment Similarity Weight)를 나타내며, 감성사전에서 단어의 존재 유무에 따라 적용
- 문서의 키워드가 감성사전 안에 존재할 경우 TF-IDF값과 극성 값만을 사용하고, 감성사전에 해당 단어가 없을 경우 유사도 사전을 통해 단어를 대치하여 사용하고, 유사도는 대체되는 단어의 유사도 값을 의미하며 Polarity는 단어의 극성값을 의미함

• 각각의 단어들에 대한 Sentiment Similarity Weight 산출 식

$$WSSW_w = \begin{cases} TFIDF \times Similarity \times Polarity \\ TFIDF \times Polarity \end{cases}$$

• SSW는 WSSW를 통해 Sentiment Similarity Weight 산출 식

$$SSW = \frac{1}{n} \sum_{k=1}^{n} WSSW$$

- SSW는 문서에서 추출된 명사, 동사, 형용사 키워드들의 가중치인 WSSW를 더한 값에 추출된 총 단어의 개수를 나눔. n은 단어의 개수를 의미하며, 각 문서들마다 추출되는 주요 키워드들의 개수가 다르기 때문에 균등한 값을 도출하기 위해 나누며, 문서가 포함하는 키워드의 WSSW의 평균을 계산하여 SSW(Sentiment Similarity Weight)를 구성함. SSW는 -1에서 +1까지의 범위를 가짐
- 해당 문서의 상위 키워드와 품사, 단어의 TF-IDF, Word2vec 기반 대체단어 및 유사도, 극성값, WSSW, SSW로 구성함. 각 단어의 품사에서 NNG는 일반명사, NNP는 고유명사, VV는 동사를 의미함

• Doc 15 문서에서 Sentiment Similarity Weight 도출 과정

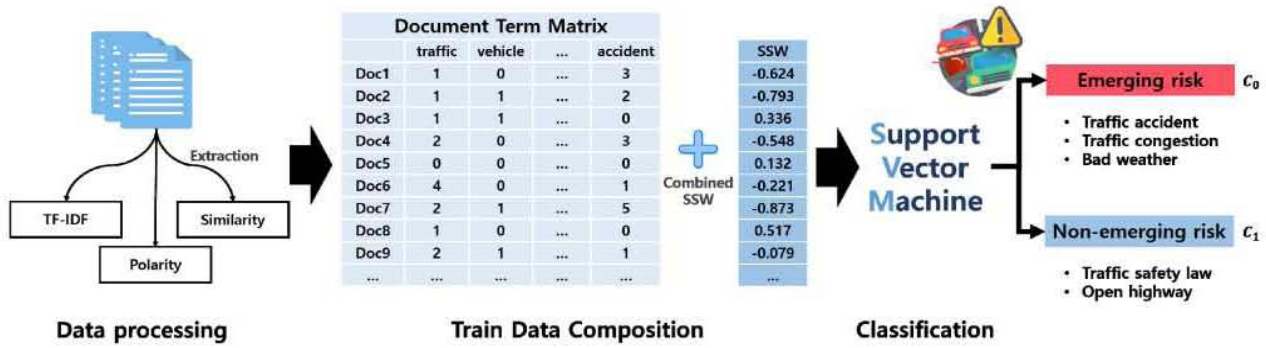
Doc 15						
Keyword	sa-go (accident)	juk-eum (dead)	JEJU	bam (night)	bi (rain)	chung-dol (crashing)
Word Class	NNG	NNG	NNP	NNG	NNG	NNG
TF-IDF	0.883	0.705	0.687	0.591	0.584	0.582
Similarity	X	X	X	X	X	crash 0.869
Polarity	−1	−1	0	+0.5	−1	−1
WSSW	−0.883	−0.705	0	0.026	−0.584	−0.505
SSW	−0.441					

- “chung-dol” (crashing)키워드는 감성사전에 존재하지 않으므로 유사도 사전에서 가장 유사한 값을 가지는 “chung-dol-ha-da” (crash)로 대체함. Sentiment Similarity Weight 도출 프로세스를 거쳐 Doc 15의 SSW가 -0.441로 산정됨

■ **감성 유사도 가중치를 이용한 교통문서 분류 모델**

- 교통 이머징 리스크 탐지를 위해 SVM 기반 분류기에 감성 유사도 가중치를 적용하여 단어의 극성값과 문서 안에서의 중요도를 고려하여 교통에 영향을 주는 이머징 리스크 관련 키워드를 감지하여 분류함
- 분류 클래스는 문서 간의 키워드 유사도를 비교하여 라벨링 하기 위해 교통사고, 교통체증 등의 교통안전과 밀접한 관련이 있는 교통문서들을 수집하고, 훈련집합, 테스트집합으로 구분하여 문서간의 유사도를 각 문서의 TF-IDF 가중치 매트릭스에 따른 코사인 유사도를 통해 비교함
- 유사 키워드를 가지고 있는 문서 간의 코사인 유사도 비교 결과 평균 0.71로 나타나기 때문에 수집한 교통문서와 0.71 이상의 코사인 유사도를 가지는 훈련, 테스트 데이터셋은 이머징 리스크인 C0으로 라벨링하고, 이머징 리스크가 아닌 C1은 고속도로 개통, 교통안전법 개정과 같은 문서 간의 비교를 통해 라벨링하며, SVM 기반 이진 분류기를 통해 이머징 리스크를 포함하는 문서 여부 판단함
- 분류 모델의 학습을 위해 교통문서들의 주요 키워드를 추출하고, 키워드에 따른 TF-IDF, 유사도, 극성값을 통해 WSSW를 계산하여 모두 더해 문서의 SSW를 도출함, 각 문서의 키워드 빈도와 SSW를 통해 SSW Matrix를 구축하고, 매트릭스는 각 문서의 키워드 빈도와 SSW를 통해 구성함
- SVM에 적용하여 분류를 수행하며, SVM 이진분류기는 특정 문서에서 이머징 리스크 존재 여부를 판단함

• 감정 유사도 가중치를 고려한 문서 분류 과정



- 각 키워드에 대한 TF-IDF, Similarity, Polarity를 통해 WSSW를 계산하고 평균을 구하여 SSW를 산출하며, 계산된 SSW를 DTM(Document Term Matrix)와 결합하여 학습 데이터를 구성함
- SSW를 고려한 훈련데이터는 SVM 이진 분류기에 학습시키며 분류 모델 학습과정에서 SSW를 고려함으로써 교통안전 관련 문서를 분류하는 척도로 작용하여 이머징 리스크를 감지함[21]

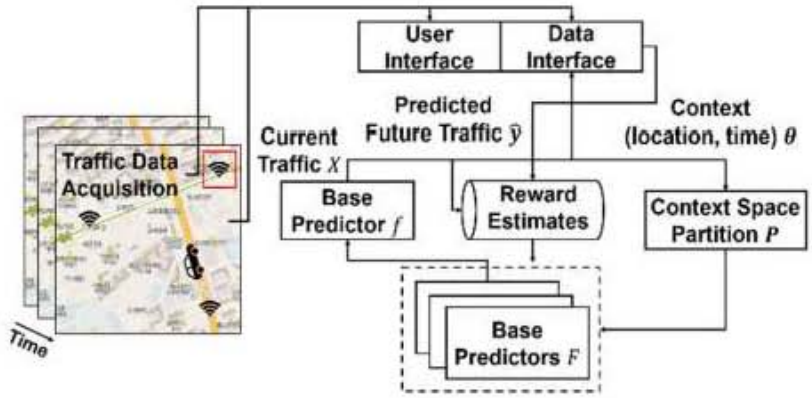
21) M. J. Kim, H. J. Kwon, J. S. Kang, J. W. Baek, K. Chung, Extraction of Road Risk Factors Based on Regression Analysis to Prevent Traffic Accidents, The 16th Asia Pacific International Conference on Information Science and Technology 2021, pp.92-94, 2021.

■ **교통 혼잡도를 이용한 사고 위험도 예측모형 개발**

- 자동차 소유의 급증으로 국내 도로 혼잡 및 교통사고를 포함한 사회 비용이 증가하고 있음. 자동차의 증가로 인한 문제는 교통 혼잡, 환경오염, 교통사고, 대중교통 불편 등의 문제점을 해결하기 위해 도입된 시스템이 ITS이며, 교통 시스템을 효율적으로 사용할 수 있는 장점을 가지고 있음
- 교통수단과 교통시설에 전자제어 및 통신 등에 첨단 기술을 사용하여 교통정보 및 서비스를 제공함. ITS는 기술의 융합을 통해 교통 체계의 운영 및 관리의 과학화, 자동화를 통해 교통의 효율성과 안정성을 향상시키는 교통 시스템이며, 실생활에서 접하고 있는 ITS는 버스 정류장의 도착 안내시스템, 교차로에서 교통량에 따른 신호처리 시스템, 내비게이션의 실시간 교통정보, 하이패스 등이 있음
- ITS는 도로나 도로변 등에 교통정보 수집 장비를 설치하여 차량 운행상황을 유선 또는 무선통신을 통해 교통관제센터로 전송하고, 교통관제센터는 수집된 정보를 교통시설에 사용하기 위해 분석하고 처리하여, 분석처리된 데이터는 교통상태를 컨트롤 하고, 사용자에게 정보전달을 함
- USC에서는 도로망의 조밀한 센서 계측 및 교통정보를 얻을 수 있는 센서(CCTV 카메라, GPS 장치 등)와 대량의 실시간 데이터 및 과거 교통 데이터를 기반으로 공간 및 시간 데이터를 구축하고 도로구간에 대한 교통 예측 문제를 해결에 활용됨

- 실시간 교통상황에서 과거데이터를 사용하여 구축된 예측 모델에 의존하는 정도는 교통 예측의 주요 과제 중 하나임. 과거의 데이터는 시간이 지나감에 따라 달라질 수 있는 한계를 극복해야 하는 문제이며, 문제를 해결하기 위해 현재의 교통상황을 실시간으로 학습하고 현재 상황을 이력 데이터로 사용하여, 이를 훈련된 가장 효과적인 예측 모델과 비교하여 교통을 예측할 수 있는 프레임워크를 개발함
- 실시간 트래픽이 도착함에 따라, 트래픽 컨텍스트 공간은 다른 상황에서 각 베이스 예측기의 효과를 효율적으로 추정하기 위해 적응적으로 분할됨
- 국내 블루시그널에서 제공하는 교통상황 예측 시스템은 단순 길 안내 서비스가 아닌 빅데이터 분석을 통한 길안내, 교통흐름 및 사고위험을 예측하는 시스템이며, 교통공학 이론을 바탕으로 인공지능 기반의 교통상황 예측 엔진을 개발함
- 기존 네비게이션 서비스는 길이 막히는 곳과 최단 시간, 최소거리 등 경로를 알려 주는 반면 블루 시그널의 예측 엔진은 가장 안전하고 편리한 주행 경로를 예측하고, 이 엔진은 국내 고속도로의 교통사고 예측은 98%의 정확도를 가짐

• USC 미디어의 교통 예측 시스템 구성도

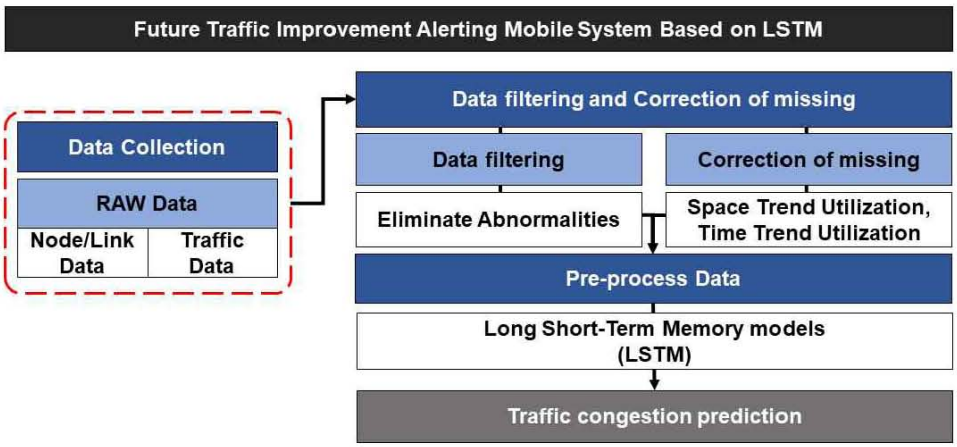


- 데이터는 사용자 인터페이스와 데이터 인터페이스에 의해 실시간으로 수신되고, Context Space의 적응적 분할을 통해 각 기본 예측 장치의 효과를 효율적으로 추정하여, 다양한 상황에서 교통상황을 예측할 수 있는 기술임

■ 시공간 결측 보정을 통한 LSTM 기반 교통 혼잡도 예측[22)]

- 교통정보시스템에서 제공하는 노드/링크 데이터와 교통 속도 데이터 수집. 노드/링크 데이터는 도로의 구간이나 도로가 이어지는 지점을 나타냄. 교통정보시스템의 교통속도 데이터는 도로나 도로변 등에서 교통정보 수집 장비를 통해 수집이 이루어짐
- 수집 장비의 오류나 도로에 차가 다니지 않아 생기는 음영지역 발생으로 인한 정보 수집 불가 및 과속으로 인한 이상치 생성이 원인이 되기 때문에 교통 데이터에는 결측 및 이상치가 포함되어 있음. 교통데이터는 시계열적인 특성을 가지는 데이터이기 때문에 이와 같은 데이터 결측은 딥러닝 모델을 통해 예측할 때 특징값 추출에 어려움이 생기므로 이상치와 결측값을 처리하여 주는 데이터 전처리를 진행함
- 데이터 전처리는 이상치부터 처리하며, 중위절대편차 방법을 통해 필터링 작업 진행. 데이터 결측 보정은 공간추세 활용법과 시간추세 활용법 및 패턴 활용법을 통해 보정작업 진행. 전처리가 완료된 데이터를 가지고 LSTM을 활용하여 교통 혼잡도 예측

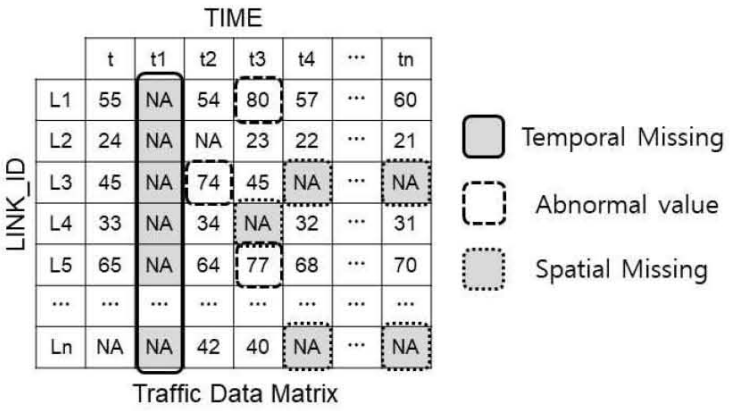
• 시공간 결측 보정을 통한 LSTM 기반 교통 혼잡도를 예측 프로세스



▪ 교통 데이터 특성에 따른 이상치 및 결측 유형과 보정 방법

- 교통 속도 데이터에는 교통 평균속도의 흐름을 왜곡시키는 이상치와 수집이 되지 않은 데이터 결측이 존재함. 이상치는 도로마다 교통흐름의 평균이 존재하지만 교통흐름의 평균에 벗어나 지나치게 작거나 큰 값을 나타내며, 예측을 진행할 때 특징 값에 영향을 끼치므로 제거가 필요함
- 교통 데이터가 5분마다 수집되어야 하지만 5분에 해당되는 전체 데이터가 수집되지 않는 경우는 시간적 결측. 또한 5분마다 수집은 되었지만 도로 별로 수집이 되지 않는 부분이 있는 경우는 공간적 결측

• 이상치와 결측이 발생되는 유형에 대해 시간별, 링크 별 매트릭스



- 교통데이터의 이상치 및 결측의 유형을 보정하기 위해, 이상치 제거 과정부터 진행하고, 이상치 제거의 대

22) D. H. Shin, K. Chung, R. C. Park (2020) Prediction of Traffic Congestion Based on LSTM Through Correction of Missing Temporal and Spatial Data, IEEE Access, 8, 150784-150796.

표적인 방법은 중위 절대편차, 중앙값, 절사평균, 변이 추정 등 탐색적 데이터 전처리 방법들이 존재함. 각 도로의 특성과 교통 데이터 특성에 따라 적절한 방법을 사용하거나 반복적인 평가에 따라 조합할 수 있음

- 이상치 제거를 위해 절대 편차의 한 종류로서 수집된 데이터의 중앙값을 사용하여 비정상적으로 크거나 작은 값들을 탐지하는 중위 절대 편차 방법을 사용함. 탐지된 값들은 이상치로 간주하여 제거함

• 이상치 제거 알고리즘

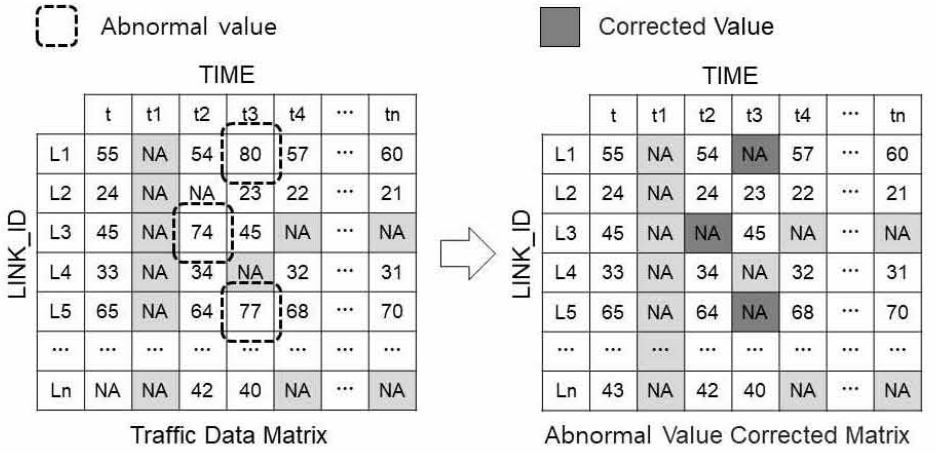
Algorithm 1 Outlier Removal Algorithm

Input: $[x_1, x_2, \ldots, x_n]$

def *Detection of Outlier*
  $MED \leftarrow Median([x_1, x_2, \ldots, x_n])$
  for x_i *in* $[x_1, x_2, \ldots, x_n]$
    do $x_i' = |x_i - median|$
  $MAD \leftarrow Mean[x_1', x_2', \ldots, x_n']$
  for x_i' *in* $[x_1', x_2', \ldots, x_n']$
  if $\frac{0.6457(x_i - median)}{MAD} >$ ***threshold***
  then $outlier\_set \leftarrow [outlier\_set, x_i]$

Output:$[x_1, x_2, \ldots, x_n]$ - outlier set

• 중위 절대 편차를 이용한 이상치 제거 과정



	t	t1	t2	t3	t4	···	tn
L1	55	NA	54	80	57	···	60
L2	24	NA	NA	23	22	···	21
L3	45	NA	74	45	NA	···	NA
L4	33	NA	34	NA	32	···	31
L5	65	NA	64	77	68	···	70
···	···	···	···	···	···	···	···
Ln	NA	NA	42	40	NA	···	NA

	t	t1	t2	t3	t4	···	tn
L1	55	NA	54	NA	57	···	60
L2	24	NA	24	23	22	···	21
L3	45	NA	NA	45	NA	···	NA
L4	33	NA	34	NA	32	···	31
L5	65	NA	64	NA	68	···	70
···	···	···	···	···	···	···	···
Ln	43	NA	42	40	NA	···	NA

- 결측 보정과정은 필터링 과정에 의해 이상치로 판명되어 제거 처리된 데이터를 알고리즘을 통해 보정하는 과정으로, 결측 데이터 보정 방법은 교통 패턴이 유사한 구간의 데이터를 이용하는 공간추세 활용법과 과거 데이터를 통해 현재 값을 보정하는 시간 추세 활용법, 패턴 자료 이용법이 있음
- 각각의 방법은 공간과 시간의 관점을 다르게 두고 데이터에 생기는 결측값을 상호보완적으로 보정함. 먼저 공간추세 활용법은 교통 패턴이 유사한 구간, 즉 상류부의 교통의 흐름은 그대로 하류부의 교통의 흐름에 영향이 미친다는 것을 전제하에 결측을 보정함

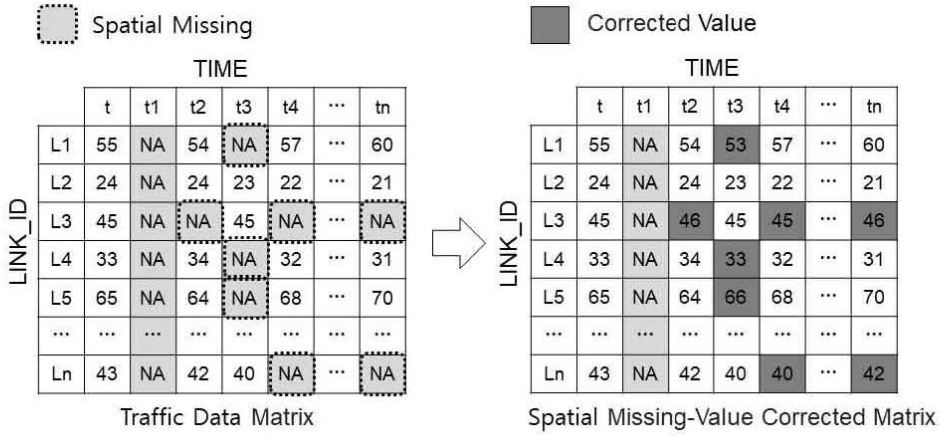
• 공간 추세 활용법을 활용한 결측 보정 과정

Algorithm 2 Spatial Data Correction Algorithm

Input: X_a (Adjoining Northbound Link) X_b (Target Link)
  X_c (Adjoining Southbound Link)
def *Spatial Data Correction*
  if $X_a = Exist$ && $X_b = None$ && $X_c = Exist$
    then $X_b = (X_a + X_c)/2$
  else if $X_a = None$ && $X_b = None$ && $X_c = Exist$
    then $X_b = X_c$
  else if $X_a = Exist$ && $X_b = None$ && $X_c = None$
    then $X_b = X_a$

Output: X_b(Target Link)

• 공간 추세 활용법을 활용한 결측 보정 과정

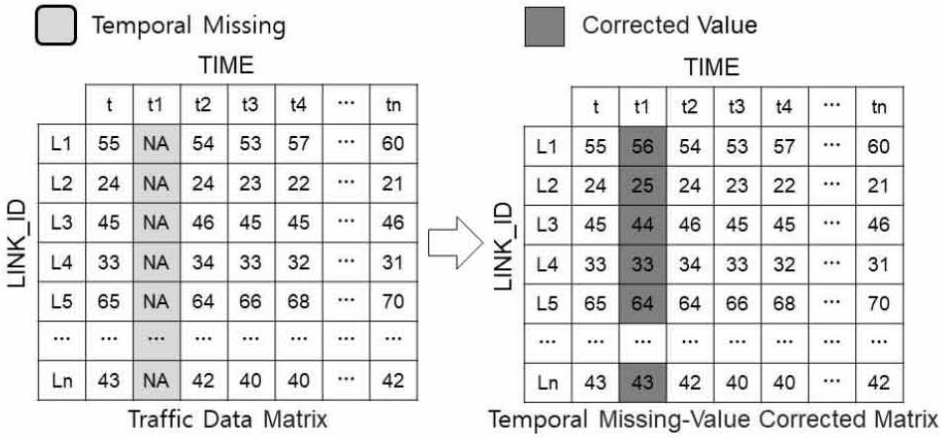


LINK_ID	t	t1	t2	t3	t4	···	tn
L1	55	NA	54	NA	57	···	60
L2	24	NA	24	23	22	···	21
L3	45	NA	NA	45	NA	···	NA
L4	33	NA	34	NA	32	···	31
L5	65	NA	64	NA	68	···	70
···	···	···	···	···	···	···	···
Ln	43	NA	42	40	NA	···	NA

LINK_ID	t	t1	t2	t3	t4	···	tn
L1	55	NA	54	53	57	···	60
L2	24	NA	24	23	22	···	21
L3	45	NA	46	45	45	···	46
L4	33	NA	34	33	32	···	31
L5	65	NA	64	66	68	···	70
···	···	···	···	···	···	···	···
Ln	43	NA	42	40	40	···	42

- 공간 추세 활용법은 연속된 3개 이상의 지점이 결측되었을 경우 인접 링크가 존재하지 않아 보정 불가. 이때 다음 방법으로 시간 추세 활용법 사용. 시간 추세 활용법은 결측이 감지된 데이터의 시점에서 이전 시점으로 n개의 자료를 평균하여 누락된 시점까지의 자료를 추정

• 시간 추세 활용법을 활용한 결측 보정 과정



LINK_ID	t	t1	t2	t3	t4	···	tn
L1	55	NA	54	53	57	···	60
L2	24	NA	24	23	22	···	21
L3	45	NA	46	45	45	···	46
L4	33	NA	34	33	32	···	31
L5	65	NA	64	66	68	···	70
···	···	···	···	···	···	···	···
Ln	43	NA	42	40	40	···	42

LINK_ID	t	t1	t2	t3	t4	···	tn
L1	55	56	54	53	57	···	60
L2	24	25	24	23	22	···	21
L3	45	44	46	45	45	···	46
L4	33	33	34	33	32	···	31
L5	65	64	64	66	68	···	70
···	···	···	···	···	···	···	···
Ln	43	43	42	40	40	···	42

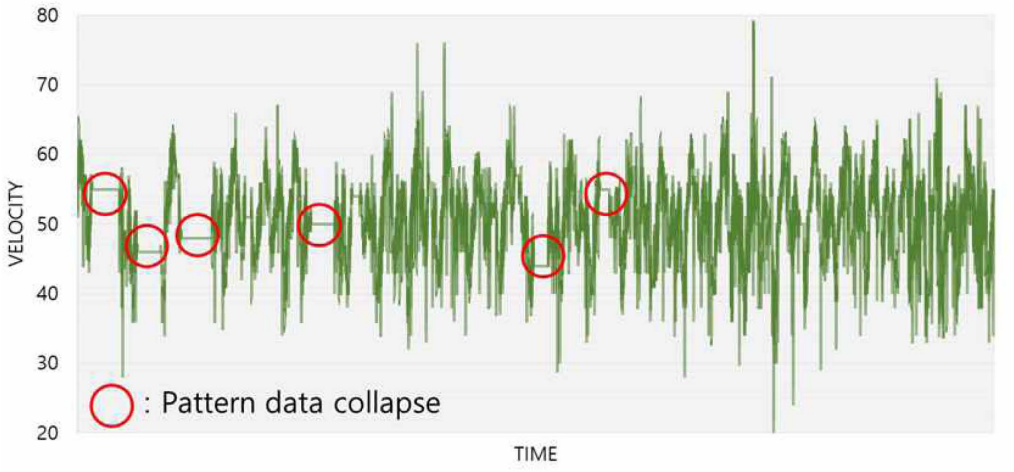
- Ft는 현재 주기 t의 누락 데이터 추정치, At-k는 기간 t-k의 검지 데이터, n은 과거 이용 데이터의 검지 주기수를 표현함
- 시간 추세 활용법을 활용하면 교통 데이터의 결측은 전부 채워지지만 시간 추세 활용법을 이용할 경우에도 결측된 지점이 연속적일 경우 시간적 결측이 연속으로 연결되어 데이터의 패턴이 사라짐

• 시간 추세 활용법의 보정식

$$F_t = \frac{A_{t-1} + A_{t-2} + \ldots + A_{t-n}}{n}$$

- 시간 추세 활용법을 사용할 수 없을 경우 패턴자료 이용법을 사용. 패턴 자료 이용법은 진출부 부분 및 진출입 접속부 부분과 같이 연결된 부분에서 수집된 자료를 이용하여 결측이 일어난 부분의 데이터를 추정할 수 있음
- 패턴데이터의 생성 절차는 과거 데이터의 요일을 확인하고 요일별 통행 특성을 파악한 다음 특수일, 일요일, 토요일, 월요일, 평일(화~목), 금요일의 6가지 유형으로 구축하여 저장함
- 유형별 패턴데이터의 생성주기는 5분이며 매 생성주기마다 현재 수집 속도에 가중치를 반영 값을 적용하여 업데이트 진행함

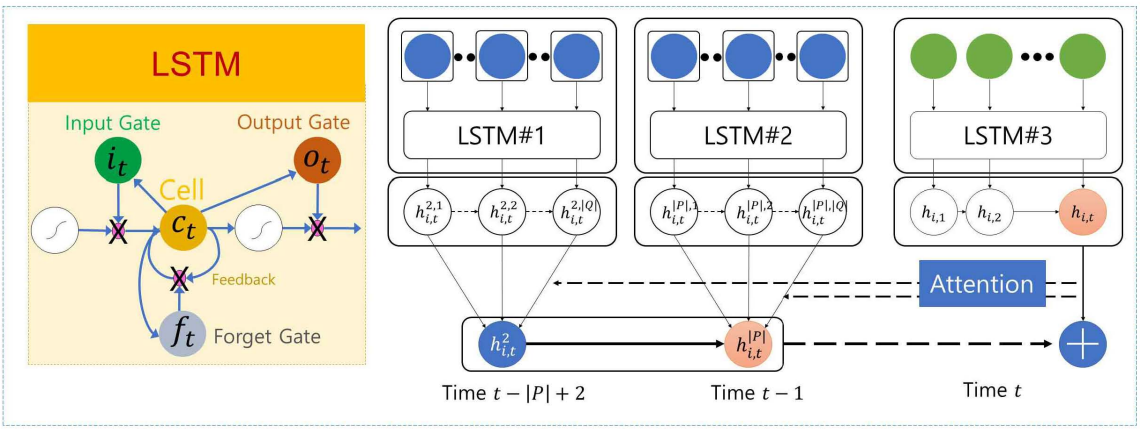
• 연속된 시간적 결측으로 인한 패턴 데이터 흐름



■ LSTM 기반 교통 혼잡도 예측 방법

- 교통속도 예측을 위해 시계열 기반의 딥러닝 모델(LSTM, Long-Term Memory)을 사용하여 모델을 구성하고, 예측을 위해 사용되는 데이터는 결측보정 방법을 통해 전처리 된 데이터를 사용함
- 모델의 구성 중 입력 데이터는 10분전 속도와 5분전 속도 평균, 현재 속도, 인접 상류부 속도 데이터를 사용. 출력 데이터는 5분 후 속도 데이터 사용함

• LSTM 기반 교통상황 예측 프로세스



- LSTM 셀의 구조는 메모리 셀과 게이트로 구성되고 메모리 셀에는 입력된 정보가 저장되고 게이트에는 셀에 저장되는 정보를 제어하기 위해 사용됨

• LSTM의 파라미터

Hyper parameter	Value
Learning rate	0.001
Dropout	0.005
Batch size	100
Optimization	RMSprop
Epoch	500

- 학습률은 예측값과 실제값의 차이를 나타내는 손실함수를 통해 한 번에 값이 이동하는 거리이며, 한 번에 너무 많은 거리를 이동하면 손실함수의 결과를 최소화하는 매개변수 W값을 지나칠 수 있으며, 너무 적은 거리를 이동하면 학습에 너무 많은 시간이 들어 효율적이지 못함
- Dropout은 학습 시 생기는 과적합을 막기 위해 사용되며, 학습 데이터에 대해서는 오류가 적지만 테스트

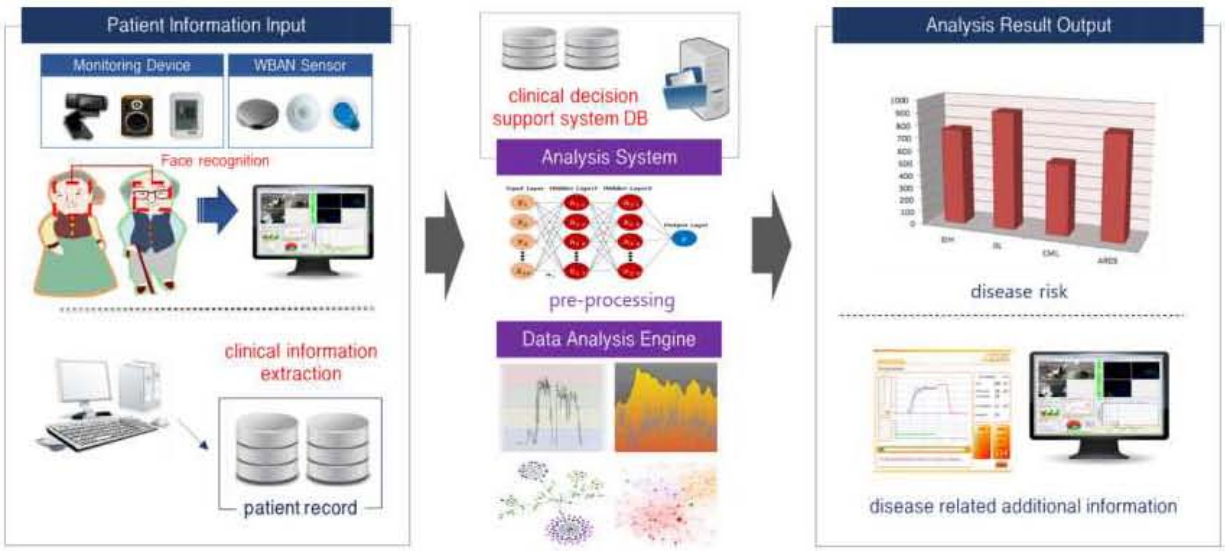
데이터로 테스트를 진행할 때 에러가 커지는 과적합 현상은 일반화가 이루어지지 않아 모델의 융통성이 부족할 때 사용됨

- Batch size는 훈련데이터 중 한번에 모델에 넘겨주는 데이터이며, optimization 함수는 가중치를 업데이트하는 알고리즘을 의미함. 성능의 결정요인이 되는 hidden neuron과 layer의 수, epoch수 loss function은 수시로 변동하여 성능 향상을 유도함

■ **운전자 안면인식 기반 사고위험도 예측 모형 개발**

- Convolution Neural Network(CNN)에서 필터링 기법을 이용한 이미지 분류 알고리즘으로, 다수의 필터와 Convolution 계층 등으로 구성된 것이 특징임. CNN의 구조는 상단 부에서 필터의 크기를 계층별로 차별하여 각기 다른 특징을 도출함
- LeNet5, GoogleNet, ResNet, DenseNet 등 다양한 CNN 학습 모델들이 파생되었으며, 이미지 분류에 월등한 성능을 보이며, 이 중 AlexNet은 ImageNet 챌린지인 LSVRC(ImageNet Lange Scale Visual Recognition Challenge)에서 8계층으로 구성된 구조로 2위를 기록하여, CNN의 우수성이 반증됨[23)]
- 한국의 과학기술정보통신부와 한국교통방송(TBN) 공동으로 개발한 교통사고 위험 예측 시스템은 경찰청과 손해보험협회를 통해 수집된 자료를 기반으로 약 6TB 규모의 빅데이터를 이용하여 해당 시스템을 학습함. 자료는 교통량, 기상, 인구, 차량 통계 등으로 200여 가지의 유의 변수를 산정하고, 학습한 시스템은 약 70% 이상의 정확도를 보임
- 국내의 Neighbor System 회사의 얼굴표정 및 생체신호 기반 라이브케어 서비스는 주변의 지인이나 친지가 없고 혼자 거주하는 노인들이 돌발 사고나 보호를 위한 사고예방 시스템으로써 사용자 위치정보, 건강정보 확인 및 종합적 상황 모니터링 서비스를 제공함

• 얼굴표정 및 생체신호 기반 라이프케어 서비스



- 안전, 건강, 가정, 응급 등 4가지의 주요 기능이 있음. 안전 기능은 사용자의 위치정보를 추적 및 관리하여 이동 경로 추적 및 이탈 감지를 통한 구인·구난 서비스를 제공함. 건강 기능은 사용자의 체온, 심박, 활동량 등을 측정 및 기록하여 이머징 헬스를 모니터링하고, 얼굴표정 분석을 통해 돌발상황이 실제 위급상황인지 아닌지를 판단하여, 이에 맞는 서비스를 제공할 수 있도록 함
- 가정 기능은 사용자의 주거공간에 설치된 CCTV를 통한 장시간 미이동, 외부침입에 대한 예방 서비스를 제공하며, 마지막으로 응급 기능은 상황 발생시 신속한 대응을 위해 기관들의 연계체계를 구축하고, 사용자의 건강 이력 등을 기관들에 전달하여 보다 신속한 대응을 함

23) ImageNet Large Scale Visual Recognition Challenge (ILSVRC), https://image-net.org/challenges/LSVRC/.

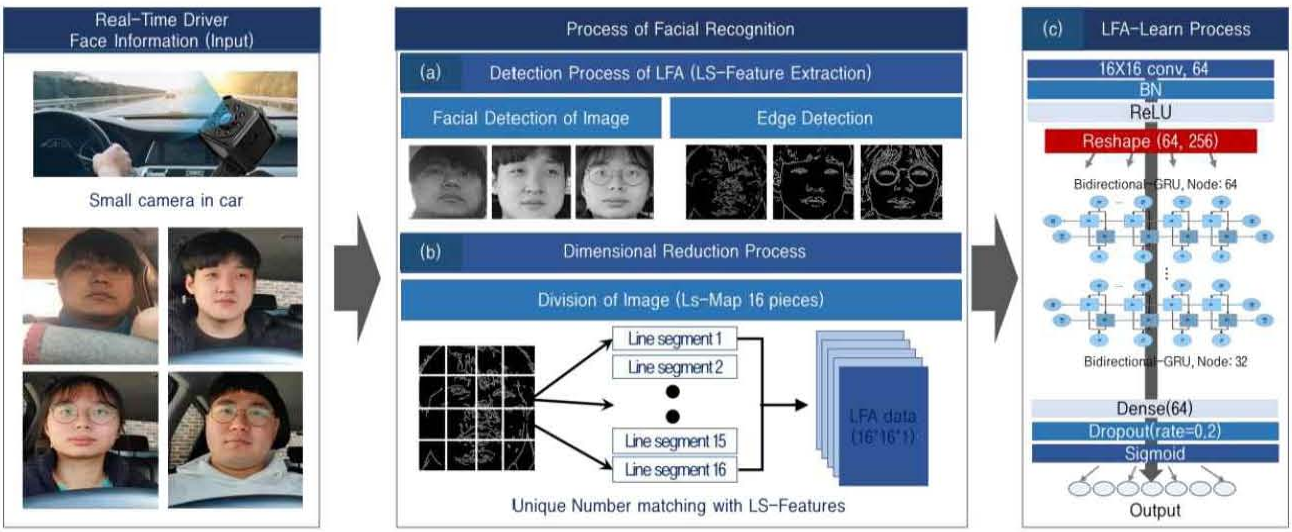
■ 위험도 예측을 위한 LFA-CRNN 기반 특징 추출을 이용한 운전자 안면인식[24]

- 고령 운전자들의 약화된 신체능력과 지각능력, 판단 능력 등의 보완 필요하며, 고령 운전자의 주행중 발생할 잠재적인 교통사고 위험과 건강 위험 상태 및 행동을 예측하여 2차 사고예방 및 건강관리와 신속한 대처 방안이 필요함. 고령 운전자들이 운전 중 급성질환으로 인해 건강상태가 악화될 경우 표정변화를 통해 사고 위험을 인지할 수 있게 됨

- 교통지원 플랫폼에서 운전자의 실시간 얼굴이미지의 특징을 추출하기 위해 Line-segment feature analysis 알고리즘과 신경망 기반 이미지 세그멘테이션 기반의 운전자의 안면을 인식할 수 있는 모델을 구성하기 위한 CRNN 개발
- 운전자 표정인식 처리 과정은 detection process, dimensionality reduction process, learn process로 구분됨. 탐지 프로세스에서 운전자의 감정 분석을 통해 상태 판단하기 위한 핵심영역(눈, 코, 입) 추출과정으로 각 영역들이 정확하게 인식되지 않은 문제를 해결함
- 핵심 영역의 특징 추출을 위해 실시간 스트림 영상에서 세분화된 프레임 형태의 얼굴 이미지에서 다중 Adaboost를 기반으로 입력된 이미지의 블록화 진행
- dimensionality reduction process 단계에서는 대용량의 원본 데이터 처리 및 고해상도에 따른 처리속도 증가를 줄이기 위해 데이터의 차원(픽셀)을 감소시켜 학습 및 추론의 시간을 감소할 수 있는 Line-Feature Analysis(LFA) 알고리즘을 이용하여, 입력 데이터의 차원(픽셀의 수)을 감소시킴

- 학습 프로세스에서 LFA 데이터를 학습할 수 있게 설계된 CRNN 모델을 통해 운전자의 표정을 인식. 운전자의 이상 상태를 파악하기 위해 UNBC-McMaster 데이터베이스를 기반으로 하여, 운전자의 상태가 고통 및 비-고통 상태인지를 판단하고 주행 중 발생될 수 있는 헬스 리스크에 대한 위험 상태를 파악 후 알림을 통해 위험상황 대처 가능

• 운전자 표정 인식 기반 헬스 리스크 추출 프로세스



■ 헬스 리스크 추출을 위한 실시간 스트림 영상 데이터 전처리 과정

- 기존 딥러닝 모델에서 얼굴인식은 입력된 얼굴 전체이미지를 활용하기 때문에, 운전자의 감정 및 아픈 상태를 분석하기 위한 핵심 요인인 눈, 코, 입 등의 영역들이 정확하게 인식되지 않는 문제가 발생
- 탐지 프로세스 모듈을 통해 차원 축소 및 학습을 수행하기 위한 전처리 작업을 진행. 실시간 스트림으로 전송되는 원본 데이터를 분석하기 위해 입력 영상을 85 frame/s로 분할하고, 표정 추출에 필요한 얼굴 이미지의 특정 부분을 멀티 블록 방식으로 추출하여 인식율 향상

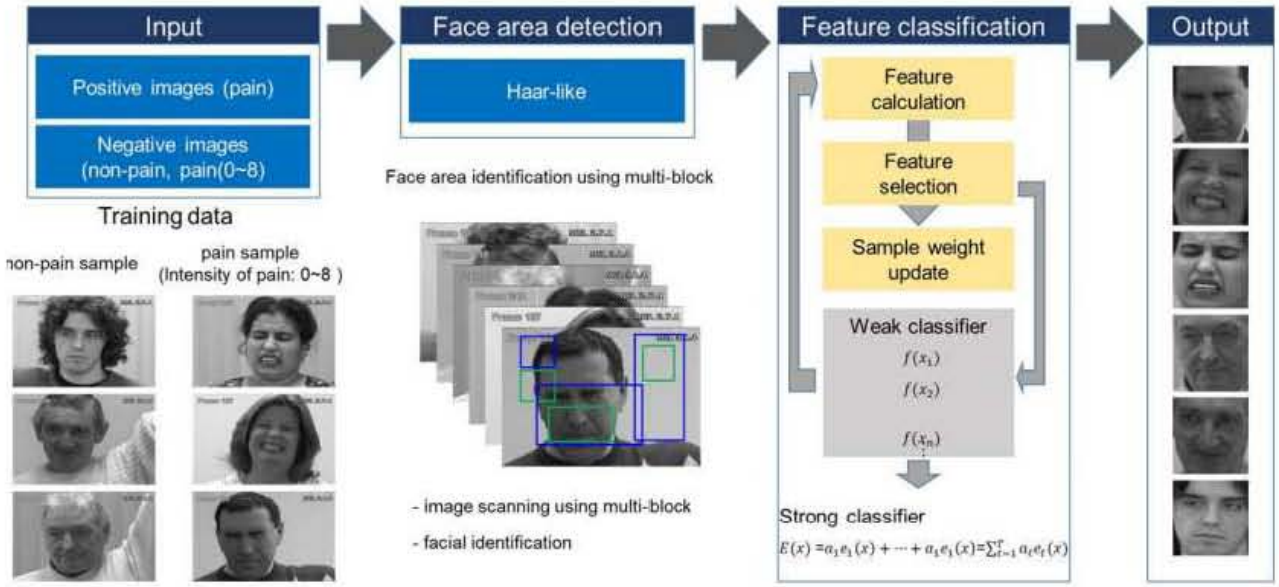
24) C. M. Kim, E. J. Hong, K. Chung, R. C. Park (2020) Driver Facial Expression Analysis Using LFA-CRNN-Based Feature Extraction for Health-Risk Decisions, Applied Sciences, 10(8), 2956-2976.

- Haar-based cascade 기반 분류기를 사용하여 얼굴을 검출하고 사용자의 얼굴에 특징점을 정확하게 추출하기 위하여 Haar-like Features를 선택, 훈련을 위해 멀티플 AdaABoost 알고리즘을 사용하고, 특징 추출은 얼굴/배경을 구분하는 특징이면서 분류기로 볼 수 있으므로 각각의 특징은 베이스 분류자와 weak classifier 후보로 정의하여 진행함

- iteration round마다 훈련데이터는 class에 맞게 분류하는 성능이 제일 좋은 단 하나의 특징을 선택하여, 해당 iteration round에서의 weak classifier 하도록 하였음
- 최종 T개의 weak classifier를 weighted linear combination 하여 최종 strong classifier를 얻을 수 있음. 이는 사용자의 얼굴에서 특징점을 검출하고 특징값의 정도에 따른 가중치를 부여하여 운전자의 아픈 감정을 판단할 수 있도록, SVM을 기반으로 각 클래스에 따라 주요 감정에 대한 가중치를 부여함

- 얼굴의 회전 정도나 얼굴의 위치에 대한 사전 정보가 없이 얼굴을 추출할 경우, 얼굴을 정규화하는 것은 매우 어려운 일이며 얼굴의 기하학적인 정보를 추출하여 얼굴을 일관성 있게 정규화를 진행해야 함

• 멀티 Adaboost 기반 얼굴의 특정 영역 검출 과정



- 얼굴의 회전 정도에 따라 분류할 수 있으며, 임의의 영상에서 사전에 이러한 정보가 없다면 얼굴 검색 시에 이러한 회전 정보를 검출함
- 얼굴 검색기의 간단한 패턴을 이용하여 멀티플 Adaboost를 통해 학습된 검출기를 직렬화하고 직렬화된 검출기를 사용하여 얼굴의 위치, 크기, 회전등의 정보를 알아낼 수 있음
- 멀티플 Adaboost 학습에 사용된 간단한 패턴은 기본적인 형태의 패턴을 이용하였으며, Adaboost 학습으로 찾아낼 간단한 검출기의 수는 160개로 선정하고 학습된 검출기는 직렬화를 통해 처리 속도 향상

▪ 실시간 스트림 영상분석의 부하 감소를 위한 LFA 알고리즘

- 얼굴 특징 추출 과정을 수행한 이후에도, 실시간 주행영상에서 운전자의 얼굴특징을 추출하기 위해서는 많은 제약조건이 발생함. 실시간 스트림 영상을 분석할 경우 영상의 흔들림으로 인해, 정확한 얼굴 표정 특징 추출이 어려울뿐더러, 입력되는 영상데이터의 용량이 크기 때문에 실시간 서비스의 수행 시 리소스 사용 제한으로 인한 정확한 분석이 어렵고, 수행 불가 상태를 초래할 수 있음

- LFA 알고리즘은 대용량의 원본 데이터 처리 및 고해상도에 따른 처리속도 증가를 줄이기 위해 데이터의 차원(픽셀)을 감소시켜 학습 및 추론의 시간을 감소할 수 있는 dimensionality reduction process 과정
- 이미지에서 정보를 추출하기 위해 3 × 3 라플라시안 마스크의 파라미터를 변형한 필터에 대한 선 정보를 추출하여 이에 맞는 1차원 벡터를 생성하고, 이 벡터를 학습 모델의 입력 데이터로 활용함
- LFA는 탐지 프로세스를 통해 산출된 운전자의 얼굴-윤곽선 데이터를 이용해 선분의 종류를 파악 및 분류

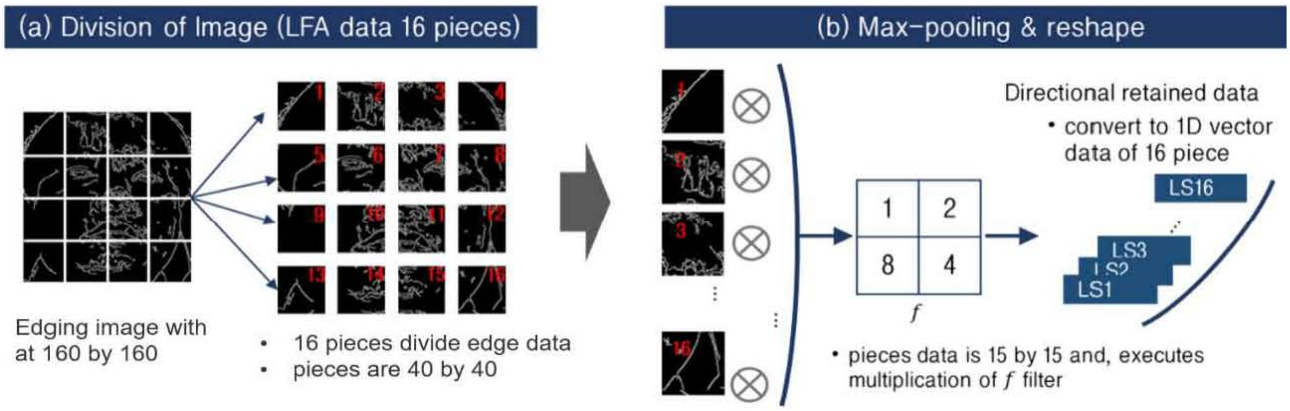
작업을 수행. 선분의 종류를 파악하기 위해 필터를 사용하며, {1, 2, 4, 8}의 원소를 가지고 있음

- 이미지 분할 알고리즘

Algorithm 1 Image Division Algorithm

Input: $[x_1, x_2, \ldots, x_n]$
def Division and Max-pooling of image
    Y = List()
    for x_i in $[x_1, x_2, \ldots, x_n]$ **do**
        sub = List()
        for w **from** 0 **to** D_w **do** // D_w, D_h denote the size of the image to be divided.
            for h **from** 0 to D_h **do**
                // fw, fh denote the size of the filter.
                sub.append(x_i[w*fw: (w + 1)*fw, h*fh: (h + 1)*fh])
        Y.append(sub)
    Y = Max-pooling(Y, stride = (2,2), padding = 'same')
Output: $Y[Y_1, Y_2, \ldots, Y_n]$

- 운전자의 얼굴-윤곽선 데이터의 분할 및 선분 종류 판단 과정



- LFA의 첫번째 처리 과정으로, 전처리 과정(detection process part)으로 산출된 윤곽선 이미지는 160 × 160의 크기를 갖고, 이 이미지를 (a)와 같이 16조각으로 분할함
- 분할된 조각은 40 × 40 크기로 이루어지며, 원본 이미지의 구성이 변환되지 않도록 조각 이미지를 배치
- 이미지를 (b)의 연산을 통해 Max-pooling 작업 후 조각들의 배치를 조정함

- Max-pooling 과정 후 조각들의 배치 조정 식

$$D_w, D_h = 4,4, P_w = \frac{W}{D_w}, P_h = \frac{H}{D_h} P[\mathrm{n},\mathrm{m}] = x[n * P_w : (n+1) * P_w, m * P_h : (m+1) * P_h],$$
$$(0 \le \mathrm{n} \le 4, 0 \le \mathrm{m} \le 4) MP[\mathrm{n},\mathrm{m}] = max - pooling(P[n,m]),$$

- 전처리 과정에서 산출된 윤곽선 이미지를 16 등분하여 "Max-pooling" 작업을 수행하는 연산 과정으로 Dw, Dh는 가로 및 세로 방향으로 분할 개수를 의미하며, Pw, Ph는 윤곽선 이미지의 크기를 Dw, Dh로 분할 시 분할된 데이터의 크기를 의미함
- P는 분할된 데이터들을 기억할 공간이며, P[n,m]를 통해 분할 위치를 유지함. n, m은 2차원 배열의 인덱스를 의미하며, 0~4의 사이 값을 갖고, MP는 분할 데이터에 대한 "Max-pooling" 함수의 결과를 기억함. 모든 과정에서는 조각(분할) 이미지들의 순서가 파괴되지 않도록 하여, 재배치된 조각들 역시 순서가 파괴되지 않도록 함

- Pain 특징 추출 데이터 처리의 부하를 감소시키기 위한 선분 집계 기반 축소 데이터 생성을 위해 실시간 스트림 영상을 기반으로 추출된 LS에 대한 선분 정보에 고유번호를 매칭

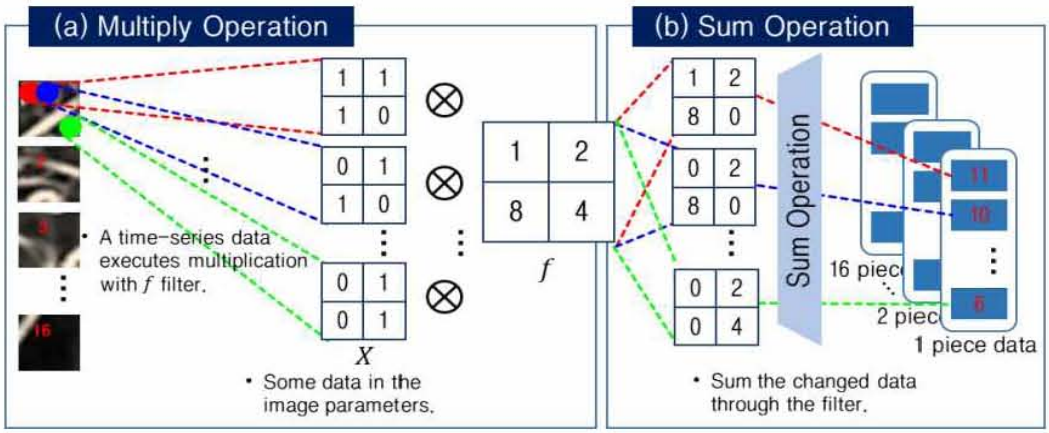
- 고유번호에서 {1, 2, 4, 8}로 서로의 값이 겹치지 않으며, 합산된 값이 서로 다른 값을 도출. LFA 알고리즘은 정상적인 선분 데이터를 매칭하기 위한 고유번호를 갖는 2×2 필터를 사용
- LS-Map은 0 또는 1인 값으로, 고유번호로 이루어진 필터가 LS-Map에 매칭될 경우 각 고유번호가 1인 영역만 반응하여 표시. 고유번호를 부여하는 이유는 시각 데이터를 이루는 모든 선분 정보를 일련의 숫자 정보로 표기하기 위함으로 곡선, 수평, 수직 등 다양한 선분들을 집계하기 쉽게 일련의 숫자 정보로 변환함. 이는 시각 데이터를 일련의 패턴(숫자) 형식으로 변환함

• 이미지를 1차원 벡터 데이터로 변환하는 알고리즘

Algorithm 2 1D Vector Conversion Algorithm

Input: $[x_1 = [p_1, p_2, \ldots, p_{16}], x_2 = [p_1, p_2, \ldots, p_{16}], \ldots, x_n = [p_1, p_2, \ldots, p_{16}]]$
def Convert image to a 1D vector
  Label = [1, 2, 8, 4]
  Y = List()
  for x_i in $[x_1, x_2, \ldots, x_n]$ **do**
  // Sub1 is a list to save the result of a piece of the image.
    sub1 = List ()
    for p_i **in** x_i **do**
    // Sub2 is a list to save the result of the image of the matched piece
    // with label data.
      sub2 = List ()
      for w **from** 0 **to** W-f_w+1 **do**
        for h **from** 0 **to** H-f_h+1 **do**
          p = p_i[w:w + f_w, h:h + f_h]
          p = p.reshape(−1) * Label
          sub2.append(sum(p))
      sub1.append(sub2)
    Y.append(sub1)
Output: $Y[Y_1, Y_2, \ldots, Y_n]$

• 이미지를 1차원 벡터 데이터로 변환하는 과정 (a) 곱하기 연산; (b) 합계 연산



- 윤곽선 데이터를 활용하여 분할된 조각 이미지는 (a)에 보이는 것과 같이 0 또는 1의 파라미터를 가짐.파라미터가 1이면 선분이며, 0일 경우 배경을 의미. 이러한 조각 데이터를 순차적으로 필터 f와 연산을 수행함
- 조각 데이터는 30 × 30 크기이며 f는 2 × 2로, 조각 데이터를 2 × 2 윈도우를 이용하여 f와 합성곱을 수행함. 윈도우는 1픽셀씩 이동(stride=1)하여 조각 이미지의 전체 영역을 스캔하고, 각 스캔 영역은 f와 연산을 통해 파라미터가 변화되어 이미지의 1 파라미터가 f 파라미터에 값으로 대체됨

• 조각 이미지와 필터 f에 대한 연산 과정

$$f = [[1,2],[8,4]] f_w, f_h = 2, 2P_i = \sum_{w=0}^{W} \sum_{h=0}^{H} x_i[w : w + f_w, h : h + f_h] \otimes f,$$

- f는 필터를 의미하며, f_w, f_h은 f의 크기를 의미하고 일정한 파라미터와 크기를 가짐. x_i는 조각 이미지의 일부 영역으로 f와 동일한 크기로 조각 이미지를 분할하여, 분할된 데이터를 f와 합성곱을 수행 후 산출된 결과를 합산하여 P_i에 기록함
- 조각 이미지의 파라미터 중 [[1, 0], [0, 1]]이 존재할 경우 필터 f와 합성곱을 수행함에 따라 [[1, 0], [0, 4]]로 조각 이미지의 파라미터가 변경되고, 변경된 파라미터들은 f로 인하여 각 파라미터 위치에 따라 각기 다른 값을 얻게 되며, 이를 합산할 경우 스캔 영역의 데이터 표현에 따라 각각의 서로 다른 결과를 도출함

• 스캔 영역에 따른 라인의 종류 및 합

Scanned Area	Summing Data	Line Type	Scanned Area	Summing Data	Line Type
0000	0	Non-Active	1000	8	Point
0001	1	Point	1001	9	Vertical
0010	2	Point	1010	10	Diagonal
0011	3	Horizontal	1011	11	Curve
0100	4	Point	1100	12	Horizontal
0101	5	Diagonal	1101	13	Curve
0110	6	Vertical	1110	14	Curve
0111	7	Curve	1111	15	Active (Side)

- 스캔 영역이 모두 "0"으로 표현될 시 해당 영역은 배경으로 간주되어, 합산 값 역시 "0"으로 표현됨. 반면 모든 영역이 "1"로 표현된 영역은 활성 영역으로, Side로 표현 "15"의 값을 얻게 됨
- 그 외에 영역들은 "1"의 위치와 개수에 따라 Point, Verticality, Horizontality, Diagonal로 표현되어 각각의 고유한 번호를 얻게 됨. 각 데이터는 동일한 선의 종류라도 표현된 위치에 따라 각기 다른 번호를 부여하며, 이의 합계를 고유 값으로 얻게 됨

■ **이미지 차원 축소를 위한 고유번호 기반 데이터 압축 및 특징맵 생성[25)]**

- 얼굴 이미지를 16개의 조각으로 분할 후 각 조각에 특정 필터를 매칭하여 산출된 정보를 통해 선의 종류에 따른 고유 값들로 구성되어 있음. 벡터 데이터들은 0~15의 파라미터들로 구성하고 고유한 특징(선분 정보)을 가짐
- 각 조각들이 가지고 있는 파라미터의 값을 기준으로 누적 집계 데이터를 생성함. 누적 집계 데이터란 파라미터의 값을 인덱스로 활용하여, 총 16 크기의 1차원 배열을 생성하고 각 인덱스가 호출될 때마다 해당 배열의 요소가 1씩 증가하는 과정을 통해 생성된 데이터를 생성함

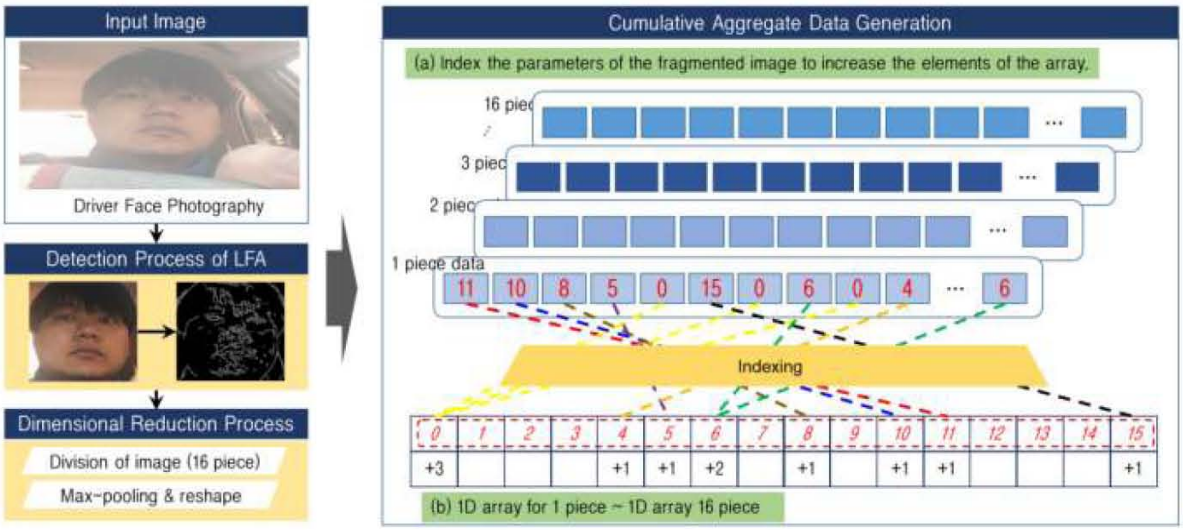
• 누적 집계 알고리즘

Algorithm 3 Cumulative Aggregation Algorithm

Input: $[x_1 = [p_1 = [v_1, v_2, \dots v_m], p_2, \dots, p_{16}], x_2, \dots, x_n]$
def Cumulative aggregation used to make LFA data
    Y = List()
    for x_i in $[x_1, x_2, \dots, x_n]$ **do**
        sub1 = List()
        for p **in** x_i **do**
            sub2 = array(16){0, ... }
            for i **from** p **do**
                sub2[i]++
            sub1.append(sub2)
        Y.append(sub1)
Output: Y

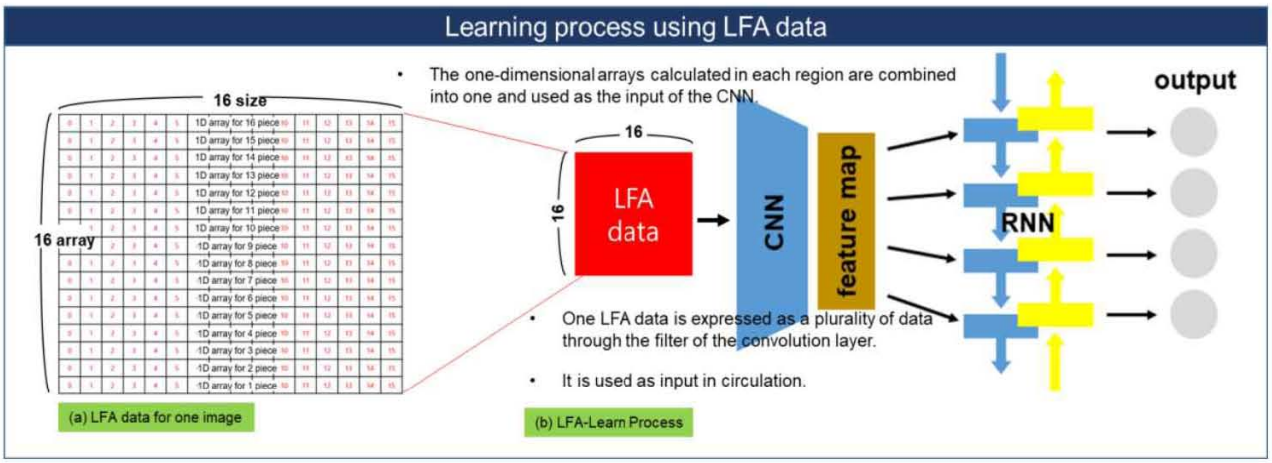
25) 특허출원, 특징 추출 기반의 LFA-CRNN을 이용한 운전자 표정 분석 장치 및 그 방법, 출원번호 10-2021-0109154, 2021.08.19.

• 누적 집계 데이터 생성 프로세스



- 조각난 데이터의 파라미터를 배열의 인덱스로 활용하여 각 조각에 따라 16 크기의 1차원 배열을 생성함. 최대 16의 크기를 갖는 1차원 배열에 각 파라미터에 상응되는 인덱스 위치에 요소의 값을 1 증가하고, 각 조각마다 수행되어 각각의 조각에 상응되는 16 크기의 배열이 생성되어 총 16개의 배열로 구성하여 LFA 데이터 구성함

• LFA 데이터를 이용한 학습 과정



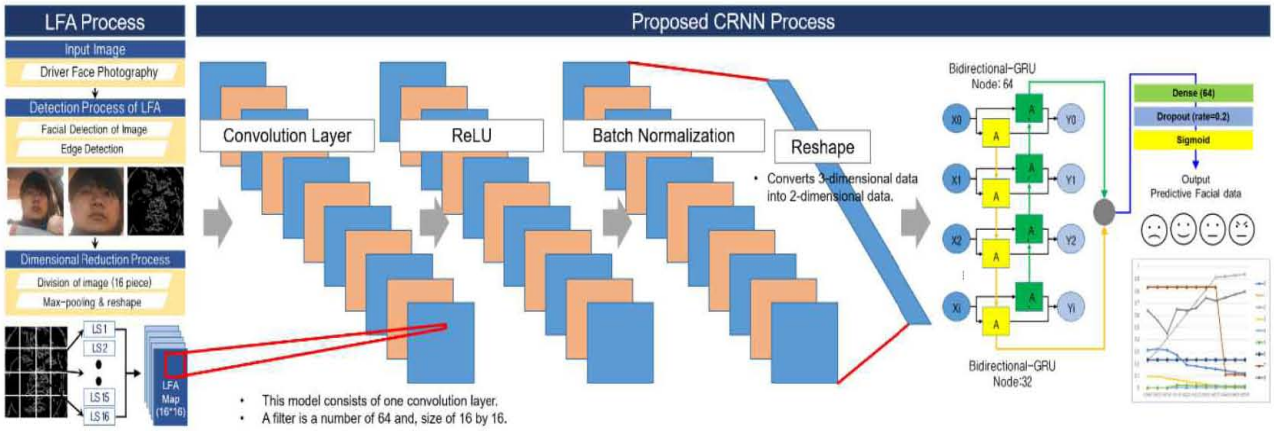
■ **운전자 상태 분석을 위한 LFA-CRNN 모델**

- 특징맵이 생성되면 주어진 입력 이미지에 대해서 전처리 과정인 얼굴 및 윤곽선 검출을 통해 도출된 이미지를 LFA 과정을 통해 16×16 크기를 갖는 2차원 배열로 재구성함. 즉, LFA 기법을 통해 차원 축소를 수행
- LFA은 항상 출력이 동일한 크기를 갖으며, 이미지가 내포한 선분에 대한 집계 정보로 구성되기 때문에 축소된 데이터 자체가 고유한 특징으로 운전자 pain 상태를 위하여 일반적인 CRNN 학습 모델의 구조가 아닌 LFA 데이터를 위한 전용 학습 모델을 설계하며, 학습 프로세스를 진행함
- LFA-CRNN 구조에서 CRNN 학습 모델로 컨볼류션 계층이 하나로 이루어지며, Reshape 계층을 통해 특징맵을 시퀀스 데이터로 표현함
- 시퀀스 데이터로 변화된 특징들은 2개의 Bidirectional-GRU(Bi-GRU)을 통해 Dense 계층으로 넘어가며, 시그모이드 계층을 마지막으로 결과 출력

- Convolution 계층의 Batch-Normalization을 통해 학습 속도 개선 및 가중치 초기값 선택의 의존성과 과적합 위험을 감소
- 차원이 축소된 LFA 데이터를 사용하기 때문에 압축된 데이터 자체가 하나의 특징이라 할 수 있음. Convolution에서 하나의 큰 특징을 여러 특징으로 표현하기 위해 16X16 크기를 갖는 총 64개의 필터를 통

해 입력에 대한 표현을 다양한 형태로 구분함

• LFA-CRNN 모델의 구조



- 도출된 값은 BN을 지나 ReLU 계층을 통해 일련의 특징맵 생성함. 이는 Reshape 계층을 통해 64개의 256 크기를 갖는 시퀀스 데이터로 재구성되며, 이를 RNN 모델의 입력으로 사용함
- RNN 모델은 2개의 BI-GRU로 구성되며 64, 32개의 노드를 가짐. 이 과정을 통해 도출된 데이터는 Dense 계층을 통해 시그모이드 함수로 전달. 이때 Dense 계층과 시그모이드 함수 사이에 Dropout 계층을 배치하여 연산량 감소 및 과적합을 방지함

- 최종적으로 시그모이드를 통해 7개의 표정 클래스를 분류. 입력인 LFA 데이터 자체가 16×16으로 상당히 작은 크기를 가지며, 이미지들에 대한 선분 누적 개수로 구성되어 해당 데이터가 압축될 경우 주요 특징들이 훼손 또는 제거할 수 있음
- 풀링 계층을 사용하지 않고 BN 및 Dropout 계층을 배치하였으며, Convolution은 stride가 "1"이며, padding은 "same"으로 설정함

■ 도로 위험 탐지를 위한 데이터 편향성 최적화 기반 연관 추론 모델[26]

- 데이터 분석 방법 중 하나인 연관분석은 데이터의 트랜잭션을 분석하여 유의미한 연관규칙을 도출하며, 연관규칙은 {X => Y}로 나타냄. 연관규칙에서 각각의 X값과 Y값은 임의의 아이템으로 구성되며, 구성된 아이템 간의 관계를 파악함으로써 의사결정에 도움을 줄 수 있음

- 연관분석에서 가장 많이 사용되는 Apriori 알고리즘은 가장 작은 아이템 집합을 기반으로 순차적으로 집합의 크기를 증가시켜 최종적으로 K개의 아이템으로 이루어진 발생 빈도가 높은 집합을 찾음. 사전에 최소지지도를 정의하며 최소지지도 이상으로 발생한 아이템 집합을 빈번 집합으로 정의하고, 정의된 n개 아이템으로 이루어진 빈번 집합은 다음 빈번 집합(n+1)의 후보 집합으로 설정되기 때문에 n개의 아이템 집합이 최소지지도를 만족하지 못하는 경우 해당 집합을 포함한 아이템 집합(n+1)은 빈발집합으로 정의가 불가함

- Apriori 알고리즘을 통해 도출된 연관분석 결과를 평가하는 지표로 신뢰도, 향상도, 지지도 등이 존재하며, 해당 평가지표를 통해 전체 데이터셋 내의 아이템의 비율과 연관도를 도출함으로써 생성된 Apriori 규칙의 타당성을 뒷받침함
- 데이터 내의 편향된 값이 존재할 경우, Apriori 알고리즘은 아이템의 발생 빈도를 기반으로 탐색하는 기법이기 때문에 무의미한 규칙들이 도출될 수 있음. 이는 데이터의 편향성을 고려함으로써 연관분석을 진행하여 한계점을 해결함

26) 류성은, 김현진, 구병국, 권혜정, 박찬홍, 정경용 (2020) 도로 위험 탐지를 위한 데이터 편향성 최적화 기반 연관 추론 모델, 한국융합학회논문지, 11(9), 1-6.

■ **교통사고 정보 데이터 수집 및 전처리**[27]

- 도로 위험 탐지를 위한 모델을 구현하기 위해 교통사고가 발생한 시점, 사고의 유형과 교통사고 당시의 주변 환경과 같은 교통사고 정보 데이터의 수집(도로교통공단 제공한 데이터 수집 및 전처리 작업을 진행)

• 교통사고 정보 트랜잭션

No.	Items
1	night, sat, car_to_car, fron_collision, no_safe_duty, single_road, offender_car, victim_car, caution
2	day, mon, alone_car, abalone, no_safe_duty, cross_road, offender_van, victim_no, caution
3	night, mon, car_to_car, lat_collision, over_speed, cross_road, offender_van, victim_cycle, warning
...	...

- 교통사고의 발생 날짜 및 시간, 부상자의 유형 및 수 등에 대한 정보로 트랜잭션을 구성함
- 교통사고 부상자, 중상자, 사망자의 수에 따라서 각각의 트랜잭션 아이템으로 설정할 경우, 트랜잭션 간의 중복되는 아이템이 적어지기 때문에 rhs(right hand side)에 교통사고의 정도가 표시될 확률이 감소함
- TAAS(Traffic Accident Analysis System)[28]에 표시된 객관적 척도에 따라 도로 위험도 지수 산출
- 도로 위험도 지수 산출을 위해 사망자, 중상자, 경상자, 부상자 수 정보를 이용하여 주의/위험/심각의 3가지 범주 형태로 분류하고, 한 번의 교통사고에 대한 정보를 나타내는 각 행의 데이터를 하나의 트랜잭션으로 구성하여 트랜잭션 데이터를 생성

■ **교통사고 정보를 이용한 연관패턴 생성**

- 교통사고 정보를 이용한 연관패턴 생성을 위해 연관규칙 마이닝 진행. 해당 데이터는 범주형 데이터로 주의-2,729개, 위험-441개, 심각-63개로 분류되어 있어 데이터 편향성이 존재
- 편향성을 감소시키기 위해서 연관규칙에서 지지도의 값을 0.02의 값으로 Prunning 단계를 진행함
- 신뢰도는 lhs(left hand side) 아이템 X가 트랜잭션에 있을 때, 아이템 Y가 존재할 확률을 나타내기 때문에 유의미한 규칙 연관성을 추출하기 위해 중요함. 신뢰도 값은 0.1로 가지치기 진행
- 마지막으로 향상도를 통해서 상위 50개의 규칙을 출력함

• 추출한 연관규칙의 상위 3개의 향상도 값을 바탕으로 도출된 규칙

Rank	Rule	Supp.	Conf.	Lift
1	{cross_road, no_safe_duty, night, car_to_man} => {caution}	0.0216	0.186	1.364
2	{single_road, over_speed, day, offender_van} => {caution}	0.0213	0.184	1.349
3	{offender_van, over_speed, night, victim_pass} => {caution}	0.0207	0.181	1.328

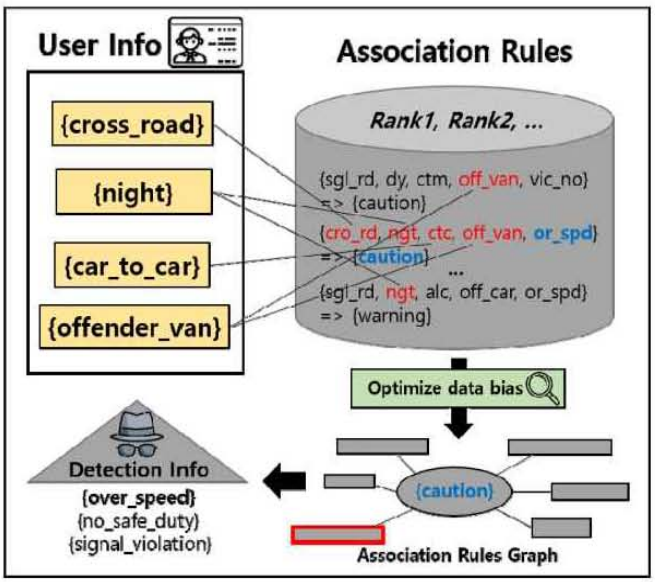
- 연관규칙들은 어떠한 아이템 조합으로도 규칙이 생성되고, lhs에 설명변수가 발생하면 종속변수 rhs가 발생하는 규칙만을 추출해야 함. 아이템 Y에는 주의, 위험, 심각에 해당하는 값들만 추출하도록 제한하며, Rank 1에 해당하는 규칙을 보면 cross_road, no_safe_duty, night, car_to_man의 상황에서 주의를 나타내는 caution 아이템이 추출되는 것을 알 수 있음

27) Traffic Accident Information, https://www.data.go.kr/data/15003493/fileData.do
28) Hazardous Road Forecasting System http://taas.koroad.or.kr/.

■ **데이터 편향성을 고려한 연관 마이닝 기반 도로 위험 탐지 모델**[29]

- 도로 위험 탐지를 위한 연관규칙 마이닝 기반의 교통 모델은 트랜잭션에서 의미있는 연관관계를 파악함으로써 기존 교통사고 정보 데이터에서 나타나지 않은 숨겨진 관계들을 도출
- 연관분석 후 도출된 관계들을 통해 도로 위험성을 탐지하며, 사용자의 개인 특성이나 주변 상황 및 환경에 따라 변화함
- 특정 환경이 주어진 사용자는 주의나 경고, 심각 등 총 3단계에 해당하는 정보와 행동 지침을 제공함
- 교통 모델을 통해 사용자의 특정 환경에 맞는 위험 탐지가 가능. 사용자에 따라 상이한 상황정보에 따라 생성된 규칙들을 통해 사용자 개인에게 적합한 정보를 제공함으로써 여러 종류의 사고를 능동적으로 예방 가능함

• 사용자를 위한 탐지 정보제공 프로세스



- 사용자가 다양한 상황을 고려하도록 구성하여 rhs에 해당하는 변수의 편향성을 감소시키기 위해 신뢰도와 지지도를 최적화함으로써, 연관규칙 그래프를 생성하여 사용자가 안전 주행할 수 있도록 데이터를 탐지하여 제공함

■ **사용자별 라이프로그 기반 헬스정보 분석 모델 개발**

- IoT, 웨어러블과 같은 스마트 디바이스의 수요가 증가하고 언제 어디서나 인터넷 사용이 활성화되고 있으며, 사용자는 IoT 디바이스를 이용하여 SNS으로부터 분산되어 있는 방대한 양의 데이터를 실시간으로 수집 및 공유와 의사소통이 가능해짐에 따라 방대한 양의 데이터를 수집 및 처리하는 시간을 줄일 수 있는 엣지 컴퓨팅을 사용함
- 헬스케어 기술이 고도화되면서 사용자들은 시간과 장소에 상관없이 하이브리드 P2P 네트워크에서 적절한 건강관리를 제공받을 수 있음[30]
- P2P 네트워크는 리소스를 중앙 서버 없이 독립적으로 연결된 노드 간의 공유 가능하지만 노드의 수가 증가하면 과부하 문제가 발생하게 됨. P2P에서 데이터를 전송하면 클라우드 서버에서 저장, 처리, 콘텐츠 사용 등을 처리하고, 클라우드 컴퓨팅은 하이브리드 P2P 네트워크에서 서버를 구축하기 위한 비용절감, 인력절감, 자료손실을 줄일 수 있음

- 인터넷을 사용할 수 있으면 시간과 장소에 구애받지 않고 서버에 저장된 데이터를 사용할 수 있지만, 클라우드는 서버가 해킹당하면 데이터가 유출되는 보안성 문제가 있음. 엣지 컴퓨팅은 네트워크 가장 자리에

29) 류성은, 김현진, 구병국, 권혜정, 박찬홍, 정경용 (2020) 도로 위험 탐지를 위한 데이터 편향성 최적화 기반 연관 추론 모델, 한국융합학회논문지, 11(9), 1-6.

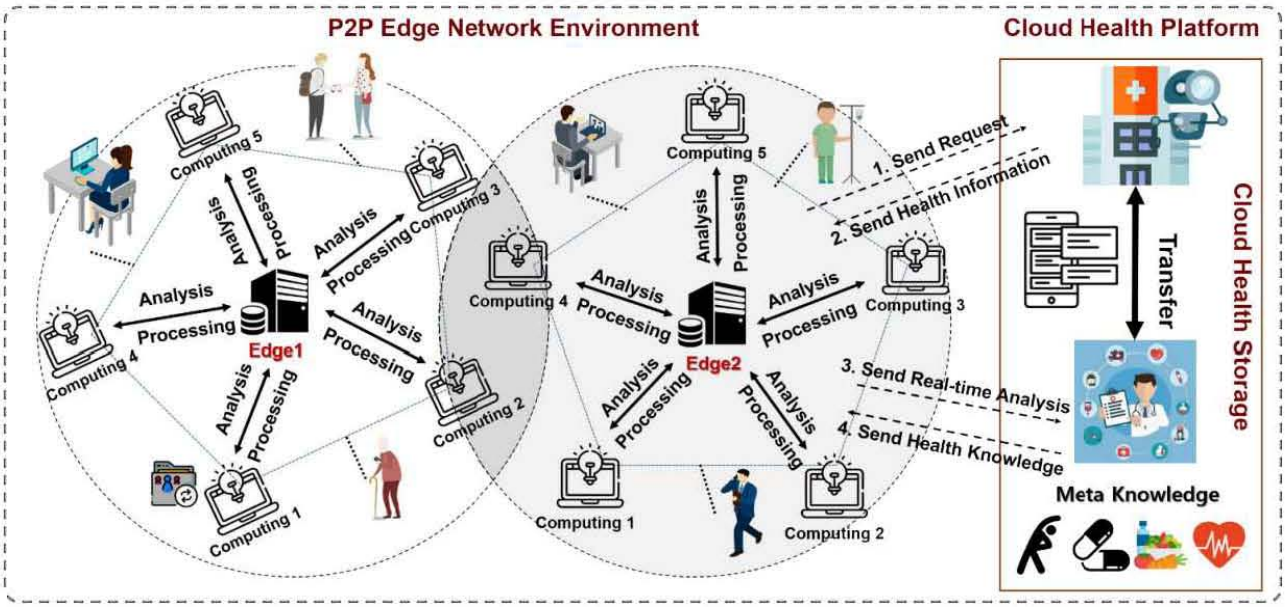
30) H. J. Kwon, J. S. Kang, J. W. Baek, K. Chung, Fast R-CNN based Imbalanced Posture Determination Model using OpenPose in Multi-Angle Images, The International Conference on Convergence Technology 2021, pp.275-276, 2021.

분산되어 존재하는 소형 서버를 통해 방대한 데이터를 실시간으로 처리할 수 있음

- 엣지 컴퓨팅은 클라우드 컴퓨팅과는 다르게 사용자들의 디바이스와 빅데이터의 연산 처리를 진행하는 컴퓨팅이 상대적으로 가까운 곳에 위치하고, 또한 분산형 시스템이기 때문에 해킹 공격의 집중 표적에서 벗어날 수 있기 때문에, 실시간으로 발생하는 데이터를 근거리에서 처리할 수 있으며, 빅데이터 처리 시간과 인터넷 사용에 필요한 대역폭 감소, 보안을 강화시킬 수 있는 장점이 있음[31]

- 하이브리드 P2P에서 개인과 개인, 개인과 의료기관이 독립적으로 연결되어 데이터를 공유 및 교환할 수 있으며, IoT 디바이스를 통해 건강 상태를 엣지 컴퓨팅의 헬스 플랫폼에서 실시간으로 처리되어 제공함

• 헬스 기반 P2P 엣지 컴퓨팅의 구성도



- 엣지 데이터에서 클라우드 저장소로 전송되기 전에 빅데이터를 처리하여 사용자에게 분석을 통해 최적의 지식을 제공함. 하이브리드 P2P 네트워킹에서 사용자가 서버에 매개변수 병렬 최적화 기법[32]으로 정보를 전송하면 서버에서 실시간으로 분석하여 처리한 후, 다시 사용자에게 정보를 피드백하여 제공함
- 헬스 서버에서 분석된 지식은 재사용과 확장을 위해 클라우드 저장소에서 메타지식으로 저장되고, 메타지식은 지식을 표현하기 위한 목적으로 사용되며 이로 인해 지식은 문제를 해결할 수 있는 의사결정 방법[33]을 제공함

31) 권혜정, 김민정, 류성은, 정경용, 클러스터링 기반 중심점 분석을 이용한 실시간 교통안전등급 산출모델, 2021 한국인터넷정보학회 춘계학술발표대회, pp. 73-74, 2021.

32) 김민정, 전병욱, 권혜정, 백지원, 정경용, BERT Bi-GRU 성능 향상을 위한 Gaussian 프로세스 기반 하이퍼파라미터 최적화, 2021 한국인터넷정보학회 추계학술발표대회, pp. 155-156, 2021.

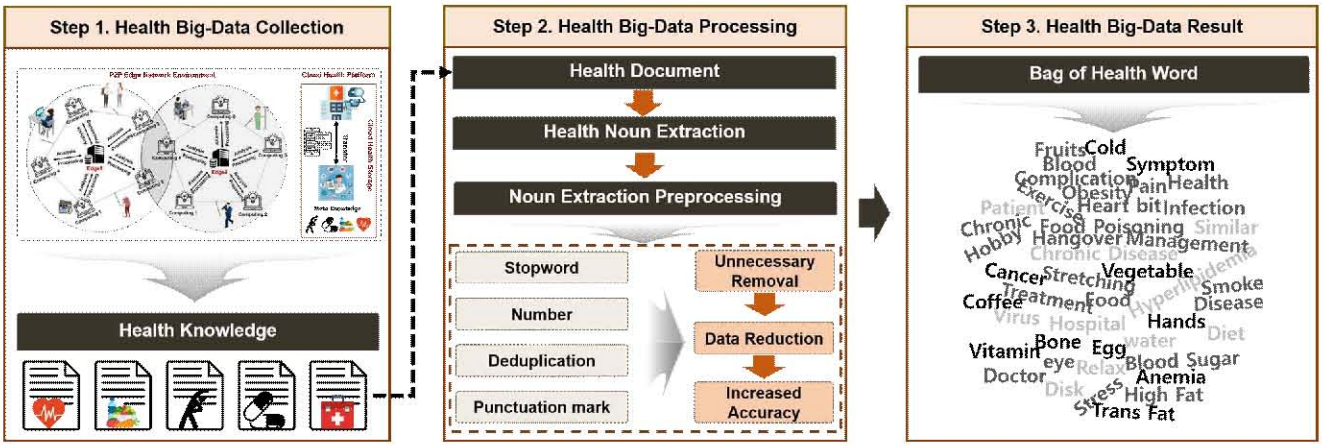
33) S. E. Ryu, H. J. Kwon, M. J. Kim, K. Chung, Deep Autoencoder Model for Unsupervised Anomaly Detection in Traffic Safety, The International Conference on Convergence Content 2020, pp. 53-54, 2020.

■ P2P 에지 네트워크에서 멀티레벨 기반 헬스 지식 마이닝 프로세스[34]

▪ P2P 에지 네트워크에서 Bag-of-health 단어 구성

- 사용자와 의료서버를 연결해 주는 헬스 플랫폼에서 하이브리드 P2P 네트워킹을 사용하여 건강관리 및 질병예방 관련 헬스 데이터를 수집함. IoT 디바이스를 이용해서 건강상태를 분석하기 위한 데이터를 헬스 플랫폼에 요청함. 엣지 네트워크를 통해 분산된 서버에서 처리되어 제공됨
- 텍스트 형식의 데이터를 전처리하여 명사 형태의 단어를 추출함. 전처리 작업은 불용어 제거, 문장부호 제거, 숫자 제거, 중복 제거 과정을 통해 필요하지 않는 요소를 제거함. 헬스 정보에서 추출한 단어로 구성된 Bag of Health Word를 구성함

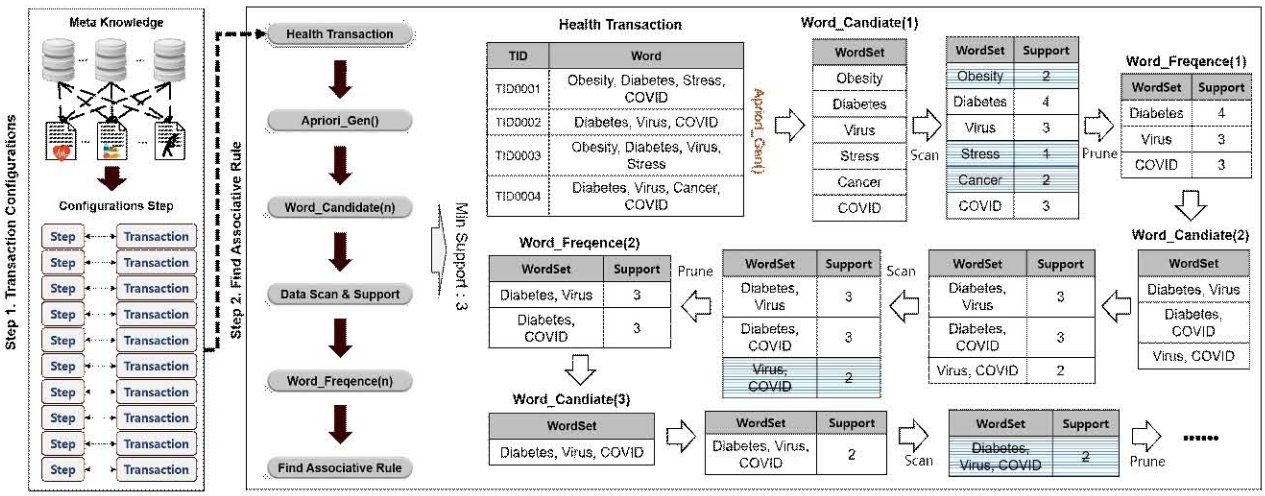
• P2P 엣지 네트워크에서 Bag of Health Word 구성 과정



▪ 지식 마이닝을 이용한 연관관계 발견

- 헬스 코퍼스에서 단어의 출현 빈도와 분포를 이용하여 헬스 트랜잭션으로 재구성함. 연관관계를 발견하기 위한 Apriori 알고리즘과 다계층 연관규칙을 사용하여 헬스 트랜잭션에서의 지식 마이닝을 진행함
- 최소지지도가 3인 트랜잭션의 빈발항목집합을 탐색하기 위해 후보항목 집합을 생성하며 Join 단계와 Prune 단계로 구성하여 후보항목집합의 수를 감소시킴. n개의 후보항목으로 구성된 빈발항목집합에서 n+1개의 후보항목집합을 생성하는 Join 단계를 진행함. 만약 n+1개의 후보항목집합의 부분집합이 트랜잭션에 존재하지 않을 경우와 최소지지도를 만족하지 못할 경우 해당 부분집합을 삭제하여 후보항목을 생성하는 Prune 단계를 진행함

• Apriori 알고리즘을 이용한 연관규칙 발견 과정



- n이 1인 A, B, C, D, E, F의 후보 항목 집합에서 최소지지도가 3인 항목을 스캔함. 최소지지도를 만족하지

34) J. W. Baek, K. Chung (2021) Multi-Level Health Knowledge Mining Process in P2P Edge Network. IEEE Access, 9, 61623-61634.

못하는 A, D, E항목을 제거하는 Prune 단계를 진행하여 n이 1인 빈발항목집합 {B}, {C}, {F}를 구성함. {B}, {C}, {F}에서 n이 2인 후보 항목 집합을 구성하는 Join 단계를 진행하여 {B, C}, {B, F}, {C, F}를 생성함

- 후보 항목 집합 생성, 최소지지도 스캔, Prune 단계 진행, 빈발항목 집합 생성, Join 단계를 반복하고, Apriori_Gen()은 트랜잭션에서 구성될 수 있는 항목 집합이 존재하지 않을 경우 빈발항목 집합 탐색을 중단하며, Word_Candidate(n)은 n개의 단어로 이루어진 후보 집합을 의미함. 생성된 후보 항목 집합을 스캔하여 최소지지도를 만족하는 집합을 정렬하여 빈발항목 집합을 생성함
- Word_Frequency(n)는 생성된 빈발항목 집합을 의미한다. 빈발항목 집합으로부터 지지도와 신뢰도를 지정하여 단어간의 연관규칙을 발견함

• 빈발항목 집합에서 추출한 연관규칙

No	Health Association Rules	Support	Confidence	Lift
1	{Obesity}→{Hypertension}	0.124	1.000	2.059
2	{High fat, Chronic Disease}→ {Trans fat, Cancer}	0.118	1.000	2.333
3	{Anemia, Diet, Cure}→ {Light head, Vitamin, Fruit}	0.100	0.875	1.494
4	{Disk, Pain, Cause}→{Position}	0.151	1.000	2.059
5	{Food, Virus}→ {Food Poisoning}	0.127	1.000	2.059
6	{Obesity, Complications}→ {Diabetes, Hyperlipidemia}	0.132	1.000	2.333
7	{Disk, Stretching}→ {Neck, Waist, Shoulder}	0.302	1.000	1.707
8	{Virus}→{Infection}	0.129	0.900	2.172
9	{Stress, Management}→ {Exercise, Sleep, Hobby, Relax}	0.129	0.900	1.537
10	{Hangover, Cure}→ {Water, Egg, Exercise, Coffee}	0.114	0.800	2.074
11	{Cancer, Ablation}→ {Operation, Radiation, Drug}	0.154	0.800	1.647
12	{Smoking, Cancer}→ { Laryngeal, Lug, Bladder}	0.129	0.900	1.537
13	{Fast-food, Overeating}→ {Cause, Complication, Obesity}	0.171	0.857	1.463
...	...	...	...	...

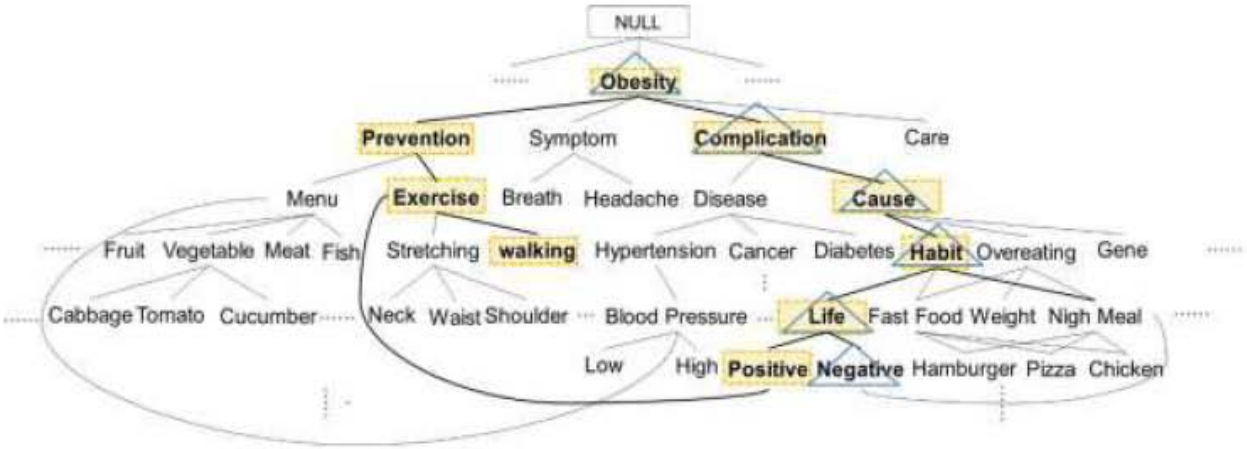
- {Obesity, Complications} => {Diabetes, Hyperlipidemia}의 연관규칙은 지지도 13.2%, 신뢰도 100%, 향상도 2.333으로 평가 척도가 구성됨. {Obesity, Complications, Diabetes, Hyperlipidemia}로 구성된 단어들간의 연관성이 매우 높고 양의 상관관계가 있음을 찾아냄

▪ 다계층 연관관계를 이용한 헬스 지식트리

- 헬스 지식트리는 다계층 연관규칙으로 구성됨. 이는 헬스 단어의 빈발항목을 구성하는 후보 항목의 연관관계와 상호정보량을 기반으로 생성함. 다계층 연관규칙은 지식을 사전에 반영하여 분류된 개념 계층을 적용하여 연관규칙을 발견하며, 개념 계층은 단어의 개념적 특징을 발견하여 단어가 가지는 의미를 계층적으로 표현함. 이는 의미를 확장하여 단어간의 유사도를 계산하고 표현함
- 다계층 연관규칙은 Apriori 알고리즘의 평가 척도인 지지도와 신뢰도를 이용하며, 후보 항목간의 관계를 발견할 수 있으며, Apriori 알고리즘 기반 깊이 탐색 방법으로 상위계층-중간계층-하위계층 형태가 만족하는 규칙을 발견하고, 데이터 항목에 대한 계층 구조를 평가하여 서로 다른 아이템에 숨겨진 패턴의 계층 레벨에 분석이 가능함
- {Obesity, Complications}의 관계를 발견하기 위해 항목 집합을 확장시켜 {Obesity, Complication, Fast Food, Chicken, Overeating, Obesity, Complications, Diabetes, Hyperlipidemia}와 같은 숨겨진 연관규칙을 발견함

- 헬스 지식트리에서 비만과 관련된 합병증의 이유로 부정적인 생활 습관을 나타내며, 생활습관 중 운동과 같은 긍정적인 습관은 비만을 예방할 수 있음을 나타내고, 계층 트리를 통해 다양한 경로 탐색으로 단어의 연관관계를 파악할 수 있음

• 다계층 연관규칙 기반 헬스 지식트리



- 연관관계의 연관정도를 파악하기 위해 헬스 지식트리를 기반으로 상호정보량을 이용한 상호연관관계를 발견함. 상호정보량(MI, Mutual Information)은 두 개의 단어가 얼마나 밀접한 관계를 맺고 있는가에 대한 상호 연관성의 정보량을 발견하는 방법이며, 단어의 각각의 빈도수의 확률과 동시에 출현하는 단어들의 빈도수 확률을 기반으로 함
- 동시출현빈도(Co-Occurrence)는 서로 다른 단어가 동시에 출현하는 횟수를 의미한다. 횟수가 큰 수일수록 두 단어 사이가 서로 밀접한 관계가 있다는 특징을 가지고 있으며, 단어간의 유사도를 측정할 때 사용하고, 정보량은 정보 이론에서 사건이 발생할 확률값을 역수에 로그를 취한값을 사용함

• 상호정보량

$$MI\,(w_1, w_2) = \log_2 \frac{f(w_1, w_2)}{f(w_1) \times f(w_2)}$$

- w1, w2는 비교할 대상을 의미한다. MI(w1,w2)은 w1과 w2사이의 상호정보량을 나타내며, f(w1)은 w1의 빈도수의 확률, f(w2)는 w2의 빈도수의 확률, f(w1,w2)는 w1과 w2이 동시에 출현하는 빈도 확률을 의미함
- 상호정보량의 결과 값이 0보다 크다면 두 단어 사이는 관련성이 있고, 0보다 작으면 관련성이 없음을 의미하며, Apriori 알고리즘에 의한 연관규칙을 평가할 수 있는 향상도(Lift)를 상호정보량에 적용함. 이는 단어간의 상호정보량을 이용하여 단어간의 연관정도를 판단하기 위함. Lift는 w1과 w2가 독립적인 관계이며 단어 w1과 단어 w2가 동시에 출현하는 빈도수를 나타내며, Lift 값에 로그를 취하면 상호정보량의 값과 동일함

• 연관 마이닝에 따른 향상도 기반 상호정보량

$$\mathrm{Lift}\,(w_1 \rightarrow w_2) = \frac{P\,(w_1 \cap w_2)}{P\,(w_1) \times P\,(w_2)}$$
$$MI\,(w_1, w_2) = \log_2 \mathrm{Lift}\,(w_1 \rightarrow w_2)$$

- {FastFood, Overeating}=>{Cause, Complication, Obesity}의 Lift 값은 1.463이며, 향상도 기반 상호정보량 공식에 대입하여 계산하게 되면 0.549가 추출되며, 상호정보량 값이 0보다 큰 값이 추출되기 때문에 {FastFood, Overeating}=>{Cause, Complication, Obesity}의 상호 연관관계가 있음을 알 수 있음

• 연관규칙에 따른 향상도 기반 상호정보량의 결과

No	Health Association Rules	Lift	MI
1	{Obesity}→{Hypertension}	2.059	1.042
2	{High fat, Chronic Disease}→ {Trans fat, Cancer}	2.333	1.222
3	{Anemia, Diet, Cure}→ {Light head, Vitamin, Fruit}	1.494	0.579
4	{Disk, Pain, Cause}→{Position}	2.059	1.042
5	{Food, Virus}→ {Food Poisoning}	2.059	1.042
6	{Obesity, Complications}→ {Diabetes, Hyperlipidemia}	2.333	1.222
7	{Disk, Stretching}→ {Neck, Waist, Shoulder}	1.707	0.772
8	{Virus}→{Infection}	2.172	1.119
9	{Stress, Management}→ {Exercise, Sleep, Hobby, Relax}	1.537	0.62
10	{Hangover, Cure}→ {Water, Egg, Exercise, Coffee}	2.074	1.052
11	{Cancer, Ablation}→ {Operation, Radiation, Drug}	1.647	0.719
12	{Smoking, Cancer}→ { Laryngeal, Lug, Bladder}	1.537	0.619
13	{Fast-food, Overeating}→ {Cause, Complication, Obesity}	1.463	0.549
....	.....		.....

- 상호정보량 수치가 어느 정도의 상관관계를 의미하는지 상관계수를 이용하여 상관 정도를 파악함. 상관계수는 크래머 C 계수를 이용하여 항목의 상관관계 발견과 의미있는 규칙에 대한 판단 기준으로 사용함. 결과값은 0과 1 사이로 표현하고, 0에 가까운 값일수록 두 관계가 연관성이 높으며, 1과 가까운 값일수록 연관성이 작은 의미가 있음

• 연관규칙에 따른 향상도 기반 상호정보량의 결과

$$C = \sqrt{\frac{\sum \frac{(O-E)^2}{E}}{N(S-1)}}$$

N : Total number of cases
O : MI(w1,w2)
B : MI(w1,w2)/sum of MI(w1,w2)
S : small number of values in row and columns

- {Hyperension, Virus, Relax, Cure, Diet, Obesity, Infection, Stress, Complications, Management}의 10개 단어의 상호정보량을 5×5 행렬로 구성하여 상관계수를 비교함
- Obesity와 Hyperension의 상호정보량은 1.042이고, 5×5 행렬로 구성하였기 때문에 N의 값은 25, S의 값은 5를 의미하고, 전체의 합은 25개의 합을 의미함. 이를 통해 상관계수 값을 추출하게 되면 0.045의 측정값이 추출되고, 1보다 0과 가까운 값이므로 Obesity와 Hyperension은 연관성이 높음을 의미함

• 단어에 따른 Cramer C 상관계수

	Obesity	Infection	Stress	Complications	…
Hypertension	0.045	0.039	0.076	0.133	…
Virus	0.066	0.058	0.033	0.088	…
Relax	0.060	0.091	0.078	0.082	…
Cure	0.070	0.087	0.088	0.092	…
Diet	0.128	0.23	0.067	0.079	…
…	…	…	…	…	…

■ **시공간 절측 보정을 통한 LSTM 기반 교통 위험도 예측 모델 개발[35]**

- 차량 이상탐지에서 이상은 램프 고장, 엔진고장 등의 차량 고장을 의미하고, 심각한 교통사고의 원인이 되기 때문에 주행 전후 또는 주행 중 운전자와 동승자, 보행자의 안전을 위해 국내외에서 차량 이상탐지에 대한 연구를 활발하게 진행하고 있음
- 기존 연구로 엔진의 진동을 감지하여 차량의 이상을 진단하는 KSDS (Knock Sensor Detection System)을 개발하여 차량에 내재된 진동감지 센서를 통해 주행 중 엔진에서 발생하는 진동신호를 실시간으로 모니터링하고, 소리의 특성을 이용하여 이상 진동 신호를 감지함
- 소리는 각각의 특성을 가지고 있으며, 파장과 진폭의 차이가 있어서 엔진의 소리 데이터를 수집하여 유형별로 특성을 분석함으로써, 소리의 특정 벡터를 추출하며 고장 부위에서 나는 소리를 인공지능 모델이 반복학습을 진행하여 고장원인을 파악함으로써, 자동차 정비사는 시간을 절약하며 자동차를 수리할 수 있고, 엔진 경고 등을 통해 운전자에게 차량고장에 대한 정보를 제공함으로써 안전하게 주행할 수 있도록 함

■ 도로 CCTV 데이터를 활용한 딥러닝 기반 차량 이상 감지[36]

■ **도로교통에서 차량의 이상을 감지를 위한 데이터 수집 및 전처리**

- 도로교통에서 차량의 이상을 감지하기 위해 Nvidia에서 주최하는 AI CITY CHALLENGE의 데이터 수집하고, 데이터 셋은 30프레임의 800 X 410 해상도의 약 15분 길이의 영상을 제공함
- 각각 100개의 학습 및 테스트 데이터를 제공하며, 학습 데이터에는 차량 이상의 지점을 영상 내 시간을 통해 주석으로 제공함
- Video_ID는 비디오 식별번호. Start_Timestamp는 이상 지점을 나타내는 시작시간(s). End_Timestamp는 이상 지점의 종료를 나타내는 시간을 나타냄

▪ 차량 이상 지점 주석

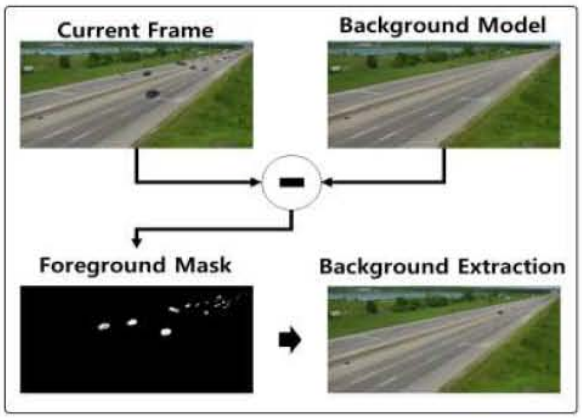
Video ID	Start Timestamp	End Timestamp
2	587	894
9	0	287
11	0	888
14	475	600
33	165	884
35	106	185
49	422	894
...	...	...

■ **MOG2를 이용한 배경 추출**

- 차량의 이상 감지를 위해 영상의 전경과 배경을 분리함. 영상에서 전경은 객체가 변위되는 부분이며 배경은 영상에서 변위되진 않는 정적인 부분을 의미함. 배경 추출을 위해 정적인 배경 모델링 진행함
- 첫 번째 단계는 초깃값을 통해 계산을 진행하여 배경을 모델링 진행함
- 두 번째 단계에서는 장면의 변화에 적응하기 위해 모델을 업데이트
- 배경 추출은 모델링 된 배경과 현재 프레임의 영상을 빼서 변위되는 부분인 전경 마스크 생성함
- 전경 마스크는 영상에서 변위되는 객체에 속하는 픽셀을 포함하는 이진 이미지를 나타냄
- 생성된 마스크를 통해 현재 프레임의 영상에서 빼주게 되면 배경 추출됨

▪ MOG2를 이용한 정지영상에서 추출된 배경

35) 특허출원, LSTM 기반 교통 혼잡도 예측 방법 및 장치, 출원번호 10-2021-0078113, 특허청, 2021.06.16.
36) 신동훈, 백지원, 박찬홍, 정경용 (2021) 도로 CCTV 데이터를 활용한 딥러닝 기반 차량 이상 감지, 한국융합학회논문지, 12(2), 1-6.



- 모델링된 배경과 현재 프레임의 이미지를 통해 프레임별 전경 마스크를 생성하고, 생성된 전경 마스크를 통해 배경을 추출. 사고 및 차량 고장의 경우 도로에서 차가 정지하며, 정지된 차는 배경 추출 시 전경에 해당하지 않으므로 배경으로 유지함

■ **COCO 데이터 집합에서 덥러닝 기반 차량 이상 감지**

- MOG2를 통해 추출된 배경에는 사고 및 고장으로 인해 정지된 차량이 포함됨. 정지된 차는 사고나 차량 고장이라는 전제를 두기 때문에 추출된 배경에 남아있는 차량을 탐지하기 위해 덥러닝 기반의 객체 탐지 모델인 YOLOv4를 사용
- YOLOv4는 기존의 덥러닝 기반 객체 탐지 모델보다 높은 FPS와 작은 mini-batch-size를 통해 적은 GPU를 통해 사용 가능. 높은 FPS를 통해 실시간 탐지를 가능하게 하며 적은 컴퓨팅 파워 사용으로 비용적인 부분 감소 가능
- COCO 데이터 셋을 통해 사전 학습된 가중치를 사용. COCO 데이터 셋[37]은 80개의 객체로 구성된 데이터이며 차량과 관련된 5개(오토바이, 승용차, 자전거, 버스, 트럭)의 객체가 포함됨. 도로 위의 차량과 관련된 객체를 탐지 가능

• 사고나 차량 고장으로 인해 정지된 이상을 탐지한 결과



- 첫 번째 배경 추출을 통해 사고나 차량 고장으로 인해 정지된 차량을 나타냄
- 두 번째 그림은 정지된 차량을 YOLOv4를 통해 이상을 탐지한 결과를 나타냄
- 파란색 테두리는 올바르게 탐지된 결과를 나타내지만 빨간색 테두리는 도로 밖의 차량이 주차되어 있는 경우로 오탐된 결과를 나타냄

■ **빅데이터 기반 교통안전 이머징 리스크 마이닝 예측**

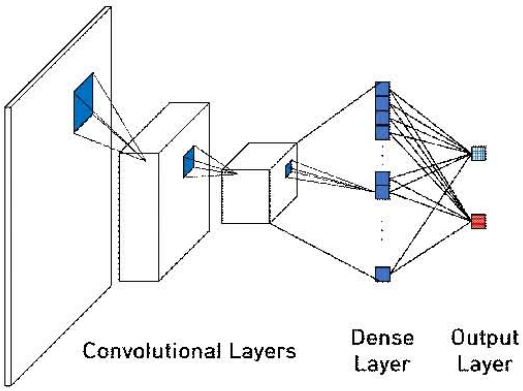
- 독립변수와 종속변수를 설정하여 데이터의 인과관계를 분석이 가능함. 분석에서 종속변수에 대해 중요도가 낮은 변수들은 분석 결과의 오류와 질을 감소시키므로 관계 분석의 질을 향상시키기 위해 중요도가 낮은 변수들을 제거함
- 독립변수들의 종속변수에 대한 기여도를 파악하는 것을 목표로 불필요한 독립변수를 제거함. 의사결정 트

37) T. Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan (2014) Microsoft coco: Common objects in context, In European conference on computer vision, 740-755.

리를 기반으로 하는 랜덤포레스트 모델에서 해당 feature가 분류의 불순도 감소에 영향을 미치는 정도를 판단하는 변수 중요도 분석 방법인 MDG(Mean Decrease Gini) 계수를 이용함

- 앙상블 기법을 사용하여 주어진 데이터를 이용하여 여러 개의 분류기를 조합하여 모델을 구성. 단일 모델들을 조합하는 방식에 따라 Bagging(Bootstrap AGGregING), Boosting, Stacking으로 구분됨
- Bagging은 샘플을 여러 번 추출하여 각 모델을 학습시켜 결과를 집계하는 Voting의 결과를 이용하며 각 샘플에서 추출된 결과를 중간값으로 조정하기 때문에 과적합 문제를 해결함. 의사결정 트리 기반 Hard-voting의 Bagging 방식을 적용한 앙상블 모델인 랜덤포레스트 학습 모델을 분류기로 적용이 가능

• 이머징 리스크 판단을 위한 CNN 분류 모델 구조[38]





(a) Sunny day road



(b) Rainy night road

- 컨볼루션 레이어는 합성 곱 연산을 수행하여 특정 영역을 강조하는 특징 맵을 출력함. 풀링 레이어는 특징 맵을 입력받은 후 크기를 줄이고 특정 데이터를 강조함. 컨볼루션 레이어와 풀링 레이어를 거쳐 입력데이터에서 특징을 추출한 후 덴스 레이어와 활성 함수 레이어를 거쳐 해당 이미지를 분류함

- AutoML(Automated Machine Learning)은 자신의 환경에 대해 스스로 학습하는 신경망 프로세스. 머신러닝 개발 과정에서 여러 알고리즘을 적용하며 데이터 셋에 최적화된 알고리즘을 자동으로 탐색함. 여러 분석 알고리즘들을 적용하여 성능이 좋은 모델들을 찾고 이들을 앙상블하여 입력 데이터에 최적화된 모델이 자동으로 도출됨[39]. 이는 모델 개발에 소요되는 시간이 감소하고 성능이 뛰어남
- 도로교통 이머징 리스크 탐지를 위한 AutoML과 CNN 기반 소프트 보팅 앙상블 분류 모델[40]을 이용하여 차량 이상 감지

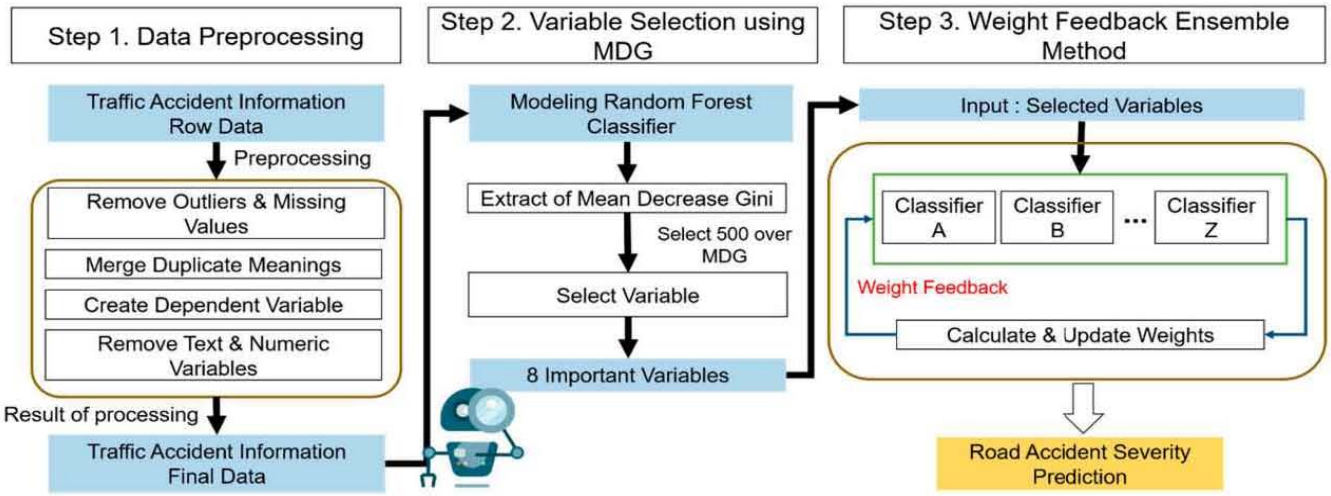
38) 권혜정, 전병욱, 김민정, 백지원, 정경용, 도로교통 이상 탐지를 위한 어텐션 마스킹 기반 적대적 학습, 2021 한국인공지능융합기술학회 추계학술대회, pp. 21-22, 2021.

39) S. E. Ryu, H. J. Kwon, M. J. Kim, K. Chung, Deep Autoencoder Model for Unsupervised Anomaly Detection in Traffic Safety, The International Conference on Convergence Content 2020, pp. 53-54, 2020.

40) 전병욱, 강지수, 정경용 (2021) 도로교통 이머징 리스크 탐지를 위한 AutoML과 CNN 기반 소프트 보팅 앙상블 분류 모델, 융합정보논문지, 11(7), 14-20.

■ 교통사고 심각도 예측을 위한 가중치 피드백 기반 Harmonic MDG-Ensemble 모델[41]

• 교통사고 심각도 예측을 위한 가중치 피드백 기반 Harmonic MDG-ensemble 모델 프로세스



■ 도로교통공단에서 제공하는 교통사고 데이터 전처리

- 주관적인 견해로 추출된 교통사고 심각도는 근거가 부족하기 때문에 신뢰도가 낮으므로 교통사고 유형에 따른 심각도를 판단할 수 있는 객관적인 판단 기준이 필요함. 교통사고 데이터 분석에서 불필요한 데이터를 제거하고 분석의 효율성을 높이기 위해서 전처리를 진행함
- 한국도로교통공단[42]의 교통사고 심각도 계산 방법에 근거하여 교통사고 심각도 변수를 생성 및 범주화함. 사망자수, 중상자수의 70%, 경상자수의 30%의 합으로 계산하여 교통사고 심각도의 값이 2 미만은 1, 2 이상 3 미만은 2, 3 이상은 3으로 범주화함

• 교통사고 정보 데이터 전처리 결과

No	Variable Name	Data Type	Category	Description
1	Traffic Accident Severity	Categorical	1 2 3	Slight-moderate Serious very serious
2	Year	Categorical	2012 2013 2014 2015 2016 2017 2018	Year of accident occurrence
12	Violation_law	Categorical	0 1 2	Pedestrian's negligence Driver's violation of law Poor maintenance of vehicle
13	Road_type	Categorical	0 1 2 3 4	Crossroad Others/Undefined One way only Parking lot Railroad crossing

- 전처리된 데이터는 17개의 변수와 18,700개의 트랜잭션으로 구성됨. 교통사고 심각도의 경우 1은 Slight, 2는 Serious, 3은 Fatal을 의미함. 법규위반은 0~2로 나타내며, 0은 보행자 과실, 1은 운전자 법규 위반, 2는 차량 정비 불량, 도로 형태의 경우 0~4까지로 나타냄. 전처리 결과만으로 심각도에 영향을 주는 요인들을 파악하기 어려움이 있으므로 심각도에 영향을 주는 중요한 변수를 추출함

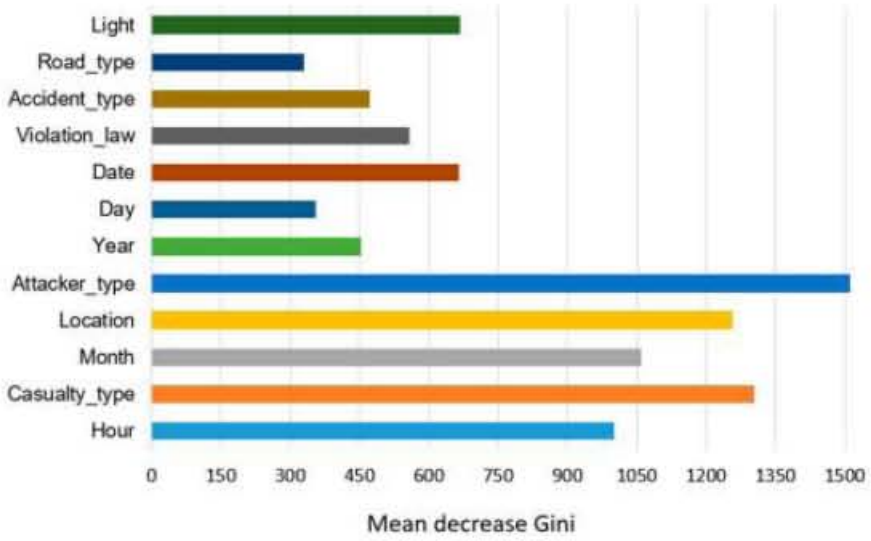
41) B. K. Koo, J. W. Baek, K. Chung (2021) Weight Feedback-Based Harmonic MDG-Ensemble Model for Prediction of Traffic Accident Severity, Applied Sciences, 11(11), 5072-5089.

42) KoROAD Traffic accident information, http://taas.koroad.or.kr/.

▪ **Mean Decrease Gini 계수를 이용한 중요도에 따른 변수 선택**

- 교통 데이터에는 심각도 예측에 불필요한 변수이거나 영향력이 약한 변수들이 존재하므로 정확한 교통사고 심각도 예측을 위해 영향력 있고 중요도가 높은 변수를 선택함. 의사결정트리 기반 랜덤포레스트를 기반으로 변수의 중요도를 도출함. MDG(Mean Decrease Gini) 계수를 이용하여 종속변수의 변화 예측에 대해 중요한 독립변수들을 선택함
- 교통사고 심각도에 대한 MDG 계수를 추출하기 위해, 교통사고 심각도를 종속변수로 설정하고 나머지 16개의 변수들은 독립변수로 설정함. 분류기를 통해 각각의 독립변수의 MDG를 비교하여 종속변수에 대한 상대적인 중요도를 판단함

• MDG 계수를 사용한 변수 중요도



- MDG 계수를 이용한 변수 중요도 결과에는 상대적으로 낮은 중요도가 추출되며 MDG 계수의 값이 500 이하일 경우에는 MDG로 선택되는 변수의 수가 급격하게 감소하기 때문에 효과적인 분석이 불가능함. 반복적인 실험을 통해 cut-off가 발생하는 임계값 설정으로 변수를 선택함

• MDG 계수를 이용한 변수 중요도의 결과

Variable Name	MDG	Variable Name	MDG
Hour	1000.54	Date	665.693849
Casualty_type	1302.854	Violation_law	557.599871
Month	1058.541	Light	667.584
Attacker_type	1512.122	Location	1256.67162

- 변수 중요도 추출 결과에서 11개의 변수가 추출됨. Hour는 MDG 계수가 1556.14로 독립변수 중 가장 큰 값이므로 교통사고 데이터 중 Hour 변수가 심각도 분류에 가장 중요한 변수라고 판단함

▪ **가중치 피드백 기반의 Harmonic MDG-ensemble 모델**

- 교통사고 심각도를 예측하기 위한 데이터를 구성하는 연속형, 범주형 등의 다범주 데이터 분류에 대한 성능을 향상시키기 위해 의사결정트리, KNN (K-Nearest Neighbor), SVM 등의 분류 알고리즘을 병합하여 앙상블 모델을 구성함
- 전처리된 교통사고 정보 데이터를 K-fold 알고리즘을 통해 훈련 집합과 테스트 집합으로 분할함. 다양한 단일 분류 모델을 학습하고 가장 성능이 우수한 분류기를 추출하여 앙상블 모델을 생성함
- 앙상블 모델의 실제 값과 예측 값을 비교하여 정답인 경우와 오답인 경우가 가장 최대가 되는 경우 모델을 선택함. 정답이 가장 큰 경우 가중치의 값을 증가시키며, 오답이 가장 큰 경우 가중치의 값을 감소시키는 조정과정을 통해 오차를 개선함

- MDG 계수로 추출된 독립변수와 정답 및 오답의 가중치를 통해 조합평균을 추출함으로써 가중치를 피드백하여 모델의 분류성능을 향상시킴. 가중치 피드백은 MDG계수에 조화평균을 적용하여 사용함. MDG 계수의 분포는 편향되는 특성이 있으므로 MDG 계수에 조화평균을 적용함
- 변수들의 중요도를 한쪽으로 치우치지 않고 균형을 유지하며 변화하는 평균 비율을 파악할 수 있는 조화평균을 변수들의 MDG 계수에 적용하여 교통사고 심각도를 예측함. 조화평균을 적용함으로써 발생하는 Harmonic MDG 계수의 값이 무한히 커지는 문제점을 정규화를 통해 해결함
- 가중치 피드백을 위해 MDG의 조화평균과 모델의 예측값의 변화량을 곱하여 본래의 가중치에 합하여 피드백함으로써 오차를 감소시키며 Harmonic MDG-ensemble 모델의 성능을 향상시킴. 앙상블 모델을 생성할 경우 발생하는 과적합 문제와 단일 모델 생성에 사용된 독립변수의 중요도를 고려하며 가중치 갱신이 가능함

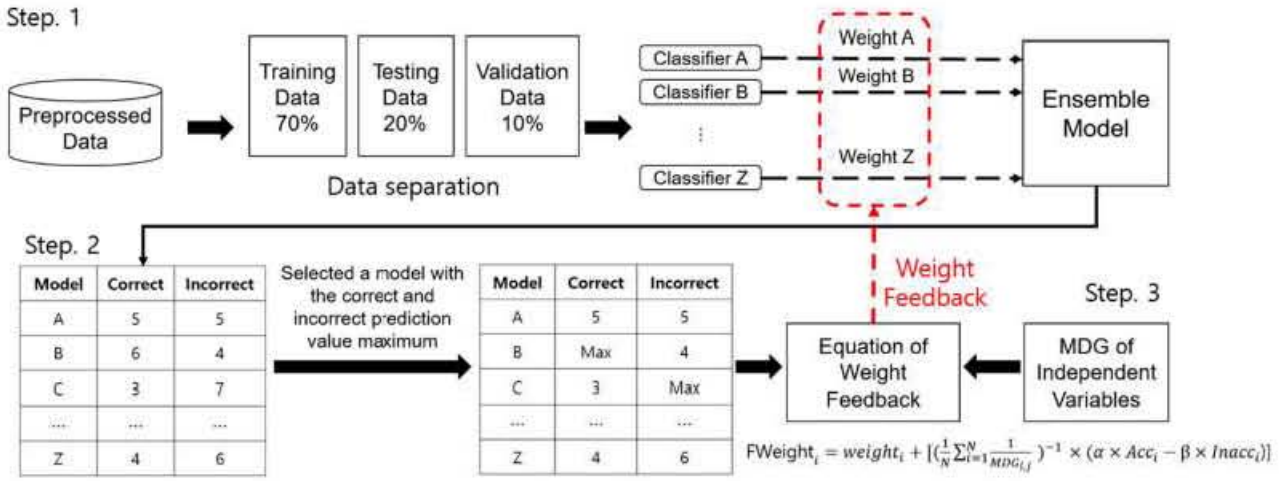
• 가중치 피드백 기반 조화 MDG 앙상블 알고리즘

Algorithm 1. Weight feedback based harmonic MDG-ensemble algorithm.

Input: Pre-processed Traffic Accident Data → TAD[j]
Output: Updated Weight using Feedback → FW
Num_Model ← Number of single models for ensemble
Correct [Num_Model] ← 0
Incorrect [Num_Model] ← 0
Accurate [Num_Model] ← 0
Inaccurate [Num_Model] ← 0
Weight [Num_Model] ← 1
Split TAD into into 70% of training data, 20% of test data, and 10% of verification data
for *k* is the number of split *TAD* **do**
  Trains data of TAD sets is number of *k*, and the rest is a dataset for testing or validation.
  Trains *Model [Num_Model]* with single classification method using train data
  Makes Ensemble model *WFEM* using *Model, Weight*
 for *j* =1 to length of train fold **do**
  Pred_e ← Prediction of WFEM for j-th transaction of test data
  Real ← Real Label of j-th transaction of test fold
  for *i* = 1 to N*um_Model* **do**
  Pred_s ← Prediction of Model[i] for j-th transaction of test data
  if (*Real* == *Pred_e*)//Under accurate prediction
  if (*Pred_e* == *Pred_s*)//Prediction of Model[i] is same as WFEM
  Correct[i]++
  else if (*Real* ! = *Pred_e*)//Under inaccurate prediction
  if (Pred_e == Pred_s)
  Incorrect[i]++
  endfor
 endfor
 Idx_Correct ← index of maximum of correct
 Idx_Incorrect ← index of maximum of incorrect
 Accurate [Idx_Correct] ← 1
 Inaccurate [Idx_Incorrect] ← 1
 $FW_i = Weight_i + H\_MDG_i(\alpha \times Acc_i - \beta \times Inacc_i)$
 Accurate ← 0
 Inaccurate ← 0
endfor
return

- 정답결과(Correct), 오답결과(Wrong)에 대한 변수를 0으로 초기화한 후 전처리된 교통사고 데이터를 k개로 분할하고 k-fold를 통해 교차검증을 진행함
- 가중치는 앙상블 모델의 예측 결과에 따라 조정을 진행하여 실제 값과 예측값의 일치 여부에 따라 기초 모델들의 Correct와 Wrong을 선택적으로 증가시킴. 반복 진행함에 따라 모든 기초 모델들의 Correct와 Wrong을 비교하여 각 값에 대해 가장 큰 모델을 선택하여 앙상블 모델을 생성함

• 가중치 피드백 기반 조화 MDG 앙상블 모델의 구조



- 가중치 증가량과 감소량을 Correct과 Wrong에 곱한 값에 MDG의 조화평균을 곱하여 기존의 가중치와 합한 결과를 피드백하여 앙상블 모델의 오차를 개선함

■ 도로교통 이머징 리스크 탐지를 위한 AutoML과 CNN 기반 소프트 보팅 앙상블 분류 모델[43]

▪ 도로교통 데이터 수집 및 전처리

- 비정형 데이터와 정형 데이터를 모두 사용하는 AutoML과 CNN 기반 앙상블 분류 모델을 설계함. CNN 기반 도로교통 이머징 리스크 분류 모델을 위하여 정상 도로의 노면과 결빙 도로의 노면 이미지 데이터를 수집함. 차량의 주행 영상이 촬영될 당시의 기상 상태를 알 수 있는 영상을 사용하여 소프트 보팅 앙상블 기반 분류함
- 결빙 도로의 노면 영상은 결빙 도로로 인해 사고가 발생한 차량의 블랙박스로 촬영된 영상을 웹에서 크롤링하여 수집함. 결빙노면과 정상노면으로 구성된 데이터를 사용함. 수집된 이미지는 노면 부분만을 크롭한 후 크기를 조정하여 224×224의 크기로 전처리함

- 기상 데이터와 사고 데이터를 수집하여 데이터를 구축함. 기상 데이터는 기상청 기상자료 개방포털을 통해 2017년부터 2019년까지 관측된 기상 데이터를 수집하며 사고 데이터는 교통사고분석 시스템에서 동기간 전국에서 발생한 사고 데이터를 수집함. 사고 데이터에서 노면 상태 정보를 추출하여 시간 및 지역 정보를 통해 기상 데이터와 합병한 후 이상치와 결측값을 제거함
- 도로 결빙 시 사고와 그 외의 사고 데이터의 비율이 약 1:48으로 불균형성으로 인한 모델의 과적합을 방지하기 위해 랜덤 언더샘플링 기법을 사용하여 두 클래스의 데이터 개수를 동일하게 구성함

▪ 앙상블 모델을 위한 CNN 기반 분류 모델과 AutoML 기반 분류 모델

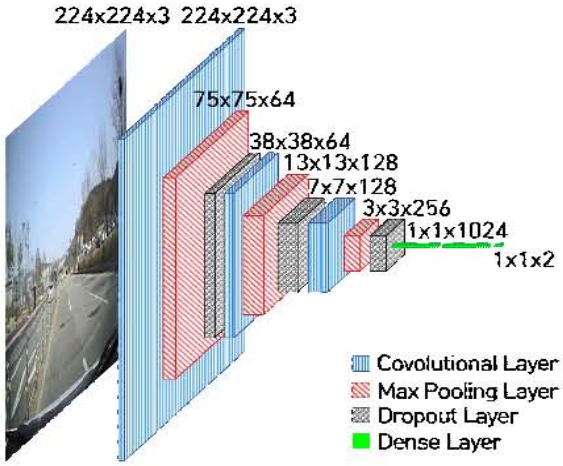
- 이미지를 통해 도로교통 이머징 리스크를 분류하기 위해 CNN 기반 분류 모델을 사용하여 특징 추출
- 모델의 구조에서 AlexNet을 기반으로 3개의 CNN 레이어와 Dense 레이어로 설계함. CNN 레이어는 컨볼루션, 풀링, 드롭아웃으로 구성되며 Dense 레이어는 2개의 1024 레이어와 2 레이어의 출력 레이어를 포함하여 설계함

- 비정형 데이터인 이미지를 입력으로 사용하는 CNN 기반 분류 모델을 활용해 도로교통 이머징 리스크의 특징을 학습하고 소프트맥스 함수를 활성 함수로 활용하여 확률값을 출력하여 데이터를 정형화함
- CNN 기반 분류 모델을 보완하기 위해 기상 데이터를 사용하여 AutoML 기반 도로교통 이머징 리스크 분류 모델을 설계함. pycaret 패키지의 AutoML을 사용하여 여러 가지의 분류 모델을 학습시키고 가장 높은 F1-score를 도출해낸 모델 3개를 앙상블하여 모델을 생성함

43) 전병욱, 강지수, 정경용 (2021) 도로교통 이머징 리스크 탐지를 위한 AutoML과 CNN 기반 소프트 보팅 앙상블 분류 모델, 융합정보논문지, 11(7), 14-20.

- Gradient Boosting classification model, Ada Boost Regression, Random Forest classification model을 앙상블하여 입력된 데이터셋에 최적화된 모델을 생성하며 도로교통 이머징 리스크 확률값을 출력함

• 도로 교통 신규 위험 감지를 위한 CNN 모델의 구조



• AutoML을 통해 학습된 분류 모델과 생성된 모델의 성능

No.	Model	Accuracy	F1-score
1	Gradient Boosting classification model	0.9022	0.9060
2	Ada Boost Regression	0.8987	0.9038
3	Random Forest classification model	0.8997	0.9023
...			
14	Quadratic Discriminant Analysis	0.7633	0.7025
Ours	**AutoML's classification model**	**0.9083**	**0.9086**

■ **소프트 보팅 앙상블 기반 도로교통 이머징 리스크 분류 모델**

- 앙상블 기법 중 결과값들을 더 유연하게 적용할 수 있는 소프트 보팅 기법을 사용함
- 앙상블 모델에 사용되는 분류 모델의 개수가 2개뿐이므로 하드 보팅에 비해 소프트 보팅이 더 적합함
- CNN 기반 분류 모델은 특징 추출을 한 후 소프트맥스 함수를 거쳐 도로교통 이머징 리스크 확률값을 출력함. AutoML 기반 분류 모델 또한 해당 이미지가 촬영되었을 당시의 기상 데이터를 입력하면 도로교통 이머징 리스크 확률값을 출력함. 가중치는 학습데이터의 이미지와 기상 데이터를 통해 도출된 확률값을 통해 새로운 AutoML 모델을 학습하여 설정함
- 분류 결과 도로교통 이머징 리스크가 예측되는 경우 운전자에게 사전에 경고함. CNN 기반 모델은 비정형 데이터인 이미지만을 입력받으나 앙상블 기법을 통해 정형 데이터인 기상 데이터를 함께 사용하는 것이 가능함

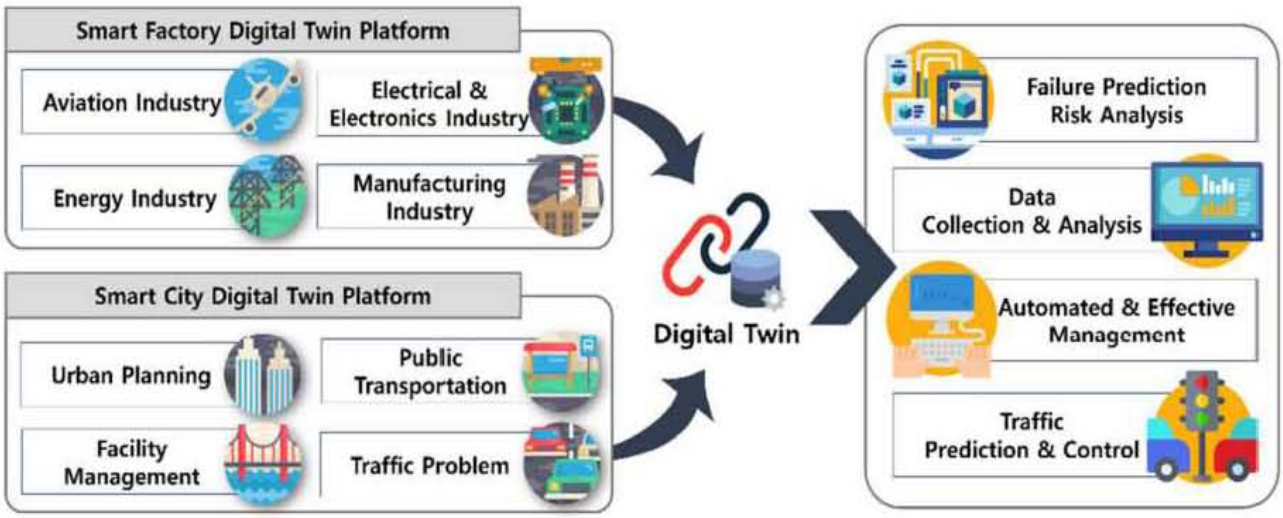
■ 빅데이터 기반 교통안전을 위한 디지털 트윈 모델[44]

- 디지털 트윈에서 현실에 존재하는 다양한 모든 것을 가상 공간에 구현하여 필요에 따라 활용 가능하기 때문에 적용 분야가 매우 넓음. 실제 데이터를 통해 가상에 구현된 디지털 트윈으로 직관적인 모니터링이 가능하고 현실과 연결된 디지털 트윈을 통해 현실의 실존하는 개체를 쉽게 제어할 수 있을 뿐만 아니라 시뮬레이션을 통한 미래 예측이 가능하여 제조, 교통, 물류, 자동차, 도시 분야 등 다양한 산업에 적용되어 문제를 해결하고 다방면으로 응용 가능할 것으로 기대되고 있음
- 디지털 트윈 구현 단계를 실제 객체 속성을 반영한 디지털 객인 Level 1, real world와 연결되어 모니터링

44) J. S. Kang, K. Chung, Ellen J. Hong (2021) Multimedia Knowledge-based Bridge Health Monitoring using Digital Twin, Multimedia Tools and Applications, 80(26), 34609-34624.

및 제어 가능한 Level 2, 인공지능의 적용을 통해 고도화된 분석 및 시뮬레이션이 활용되는 Level 3으로 구분 가능함

• 디지털 트윈의 단계별 프로세스



- 디지털 트윈 모델 생성 과정에서 실제 객체의 속성 및 다이나믹스 등을 적절하게 표현 가능한 수단이 필요함. 디지털 트윈 구현에 필요한 디지털 모델 개발에는 전통적인 데이터 기반의 모델링과 이론 및 지식기반의 시뮬레이션 모델링 방법이 사용되고 있음
- 디지털 트윈에서 사용되는 시뮬레이션 기술은 이산 사건 모델, 연속 모델, 이산 사건 모델과 연속 모델이 혼재된 하이브리드 모델로 구분되어지고, 시스템의 상태가 특정 이벤트에 의해 변화할 경우 이산 사건 모델에 해당하지만 시간에 따라 시스템의 상태가 연속적으로 변화한다면 연속 모델에 해당함
- 로봇, 자동차 등 시스템은 이산 사건 모델과 연속 모델이 혼합된 하이브리드 모델로 볼 수 있음. 하이브리드 모델은 이산 사건 모델과 연속 모델 간의 상호작용 인터페이스를 잘 정의해 주어야 함. 시뮬레이션 모델링 방법은 변수들간의 인과관계를 파악할 수 있으며, 다양한 시나리오에 대한 결과를 도출이 가능함

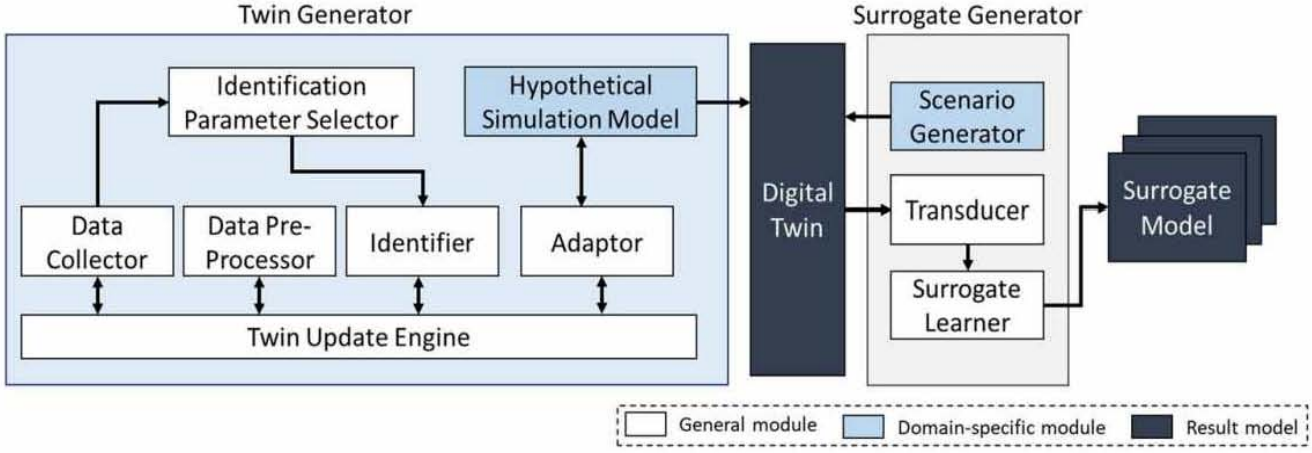
- 실제 시스템에서 수집된 적은 데이터만으로도 유의미한 검증이 가능하며, 비용이나 안전과 같은 문제로 인해 실제 시스템을 시험해볼 수 없을 경우 시뮬레이션을 통해 데이터 생성이 가능한 장점이 있음. 대상 시스템의 속성, 동작 특성, 성능 만족도 등과 같은 전문 지식을 필요로 하며 실제 시스템에서 데이터 획득이 어려울 경우, 모델의 실효성 검증에 대한 이슈가 있음
- 데이터 모델 기반 디지털 트윈은 실제 대상 시스템의 센서로부터 수집된 빅데이터와 머신러닝을 통한 학습 모델을 활용한 방법으로 시뮬레이션 모델에 비해 실행시간이 빠르고 시스템에 대한 전문 지식이 필요하지 않다는 장점이 있음. 고성능의 데이터 모델을 구축하기 위해서는 방대한 데이터가 필요하며 입출력 간의 상관관계 파악에는 유용하나 인과관계를 규명할 수 없음
- 시스템의 구성이나 운용 환경의 변경 등과 같이 학습한 조건이 변경되는 경우 이전 모델로는 예측이 어려우며 새로운 학습이 필요하다는 것이며, 데이터 수집을 위한 시간과 비용 및 학습에 상당 시간이 소요될 수 있음

■ **디지털 트윈 프레임워크 구조**

- 디지털 트윈 프레임워크에서 각 도메인에 특화된 디지털 트윈을 생성하는 트윈 생성기와 디지털 트윈 기반 제한된 시간 내에 다양한 시나리오에 대한 분석 및 실시간 서비스 제공을 위한 surrogate generator로 나누어짐
- 특정 도메인에 한정된 모듈이 아닌 Identification parameter selector module, identifier module, adaptor module은 다양한 도메인에 대해 동일하게 적용할 수 있는 부분이며 가설적 모델을 포함한 시뮬레이션 모델은 도메인 별로 특화된 기본 모델을 이용함. 가설적 모델과 입력에 따른 상태 변화를 통해 출력을 추출해내는 것은 시뮬레이션 기반 모델링 방법을 이용함으로써 가능해지고, 소량의 실데이터로부터 가설적 모

델 내부의 파라미터를 업데이트하는 것은 데이터 기반 모델링 방법을 이용함으로써 가능함

• 디지털 트윈 구조도

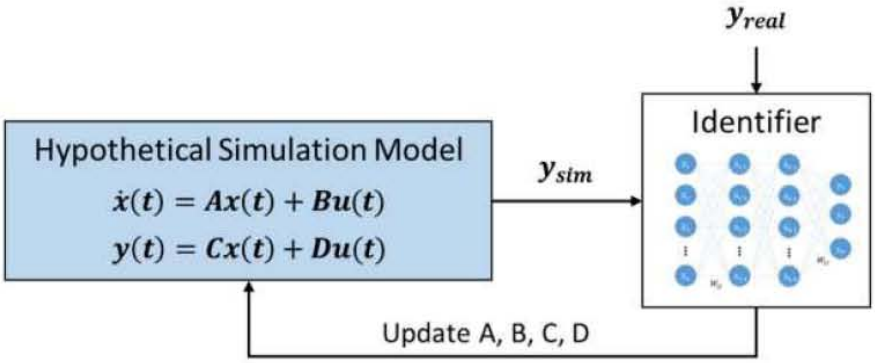


- Surrogate generator는 이렇게 완성된 twin 모델을 활용하여 실제 시스템을 대신함으로써 다양한 서비스를 제공할 수 있음. 환경 또는 상황 변화에 따라 시스템의 상태가 어떻게 변화할지 예측하여 문제점을 파악하여 사전에 예방할 수도 있으며, 시뮬레이션 모델의 특성상 다양한 시나리오를 적용하여 계획 수립에 이용할 수도 있음

■ **트윈 모델링 프로세스**

- IPS module이 가설적 시뮬레이션 모델과 데이터간의 상관관계를 파악하여 실제 시스템과 동일한 특성을 가지도록 하기 위한 적어도 하나의 고유 파라미터를 선정하는 역할을 수행함
- 모델 설계 데이터에 기초하여 구축된 실제 모델로부터 IPS module에서 선정된 대표 고유 파라미터와 관련된 데이터를 수집을 진행함. 수집된 실제 데이터를 대표 고유 파라미터와 대응하여 시뮬레이션 모델에 적용할 수 있도록 데이터 전처리 프로세서가 진행됨
- 고유 파라미터의 값은 모르지만, 입력값에 대응하는 가설 모델의 출력과 실제로 얻을 수 있는 출력과 관련된 데이터간의 상관관계를 파악할 수 있는 데이터 모델 방법을 적용한 것이 Identifier module이며, 도메인별로 다양한 형태의 시뮬레이션 모델이 가능하므로 시뮬레이션 모델의 입출력 인터페이스와 트윈 프레임워크의 생성부 내의 모듈들과의 인터페이스를 맞춰주는 역할을 수행하는 것이 adaptor 모듈임
- 모델의 내부 고유 파라미터 값을 갱신해가면서 얻은 출력값과 데이터와의 오차가 사용자가 정의한 수준을 달성하면 모델 갱신을 멈추고 실제 시스템과 유사한 동작하는 디지털 트윈이 진행됨

• Identifier 모듈의 트위닝 파라미터 업데이트 과정



■ **관계 추출을 위한 퍼지 클러스터링**

- 정보기술의 발전으로 대량의 데이터를 쉽게 수집가능하며, 다양한 분석 방법이 연구되고 있음. 데이터는 분야에 따라 내재된 정보와 형태의 차이가 존재함. 또한 정제되어 라벨링 된 데이터가 있는 반면에, 라벨링 되지 않은 데이터가 존재하기 때문에 유사관계를 파악하기 애매한 경우가 발생함. 이는 머신러닝 알고리즘

을 이용하여 데이터를 지도학습과 비지도 학습을 통해 분석 가능함

- 비지도 학습의 대표적인 알고리즘은 클러스터링이 있음. 클러스터링은 유사한 특징을 가지는 데이터를 여러 개의 그룹으로 분류하지만, 애매모호한 특징이 존재하는 데이터는 정확한 분류가 어려움. 이를 해결하기 위해서는 퍼지를 적용한 퍼지 클러스터링을 이용함
- 퍼지 관측의 소속 정도에 대한 애매모호한 정도 또는 가능성의 정도를 통해 표현 가능함. 퍼지클러스터링을 통해 애매모호한 특성을 가지는 데이터를 분류하여 관계를 파악 가능함. 퍼지 클러스터링은 관측치가 각 클러스터에 속할 가능성과 확률을 통해 분류함
- 퍼지 클러스터링을 위한 알고리즘으로 Fuzzy C-means 사용하여 클러스터 내의 관측치 간의 유사성을 최대화하고, 클러스터 간의 비유사성을 최대화하는 최적의 해를 탐색하는 방법으로 일반적인 K-means 알고리즘과 유사하지만, 각 군집에 속할 가능성과 가중치를 계산하는 과정이 존재함

■ **YOLO9000을 이용한 영상객체 추출 알고리즘**

- CNN을 기반으로 이미지 데이터에서 객체를 탐지하여 추출하는 연구가 진행되고 있음. 이를 통해 얼굴인식, 차량인식 등 다양한 분야에서 활용하고 있음. 다양한 객체 탐지 알고리즘 중에서 YOLO9000는 Darknet-19 Convolutional Neural Network를 사용하는 YOLOv2 기반 객체탐지 알고리즘이 활용됨. 이는 기존에 영상인식에서 많이 사용되고 있는 실시간 객체 검출 알고리즘인 YOLO(You Only Look Once)의 정확도를 향상시키고, WordTree 기반 이미지를 학습하여 9000개 이상의 객체 분류 및 감지가 가능함
- Batch Normalization은 neural network의 hidden layer에서 입력 분포가 학습할 때마다 변화하여 가중치의 편향성이 생기는 것을 방지하기 위한 알고리즘임. 이는 편향성을 해결하기 위해 mini batch를 하나의 단위로 하여 각 mini batch의 평균과 분산을 구하여 평균이 0, 분산이 1이 되도록 정규화를 진행함. 모든 YOLO에 사용되는 Convolution Layer에 대해 Batch Normalization을 적용하고, Dropout을 삭제하여 mAP를 향상시킬 수 있음.

- High Resolution Classifier는 모델의 학습을 위한 classifier의 input size resolution을 증가시키는 방법임. 기존 state-of-the-art는 ImageNet에서 사전에 학습된 모델을 사용하며 YOLO를 포함하여 대부분 256x256보다 작은 input size를 통해 학습함. 이는 학습을 위한 필터 조절 소요 시간이 큼. 탐지를 위해 448x448로 크기를 증가시킴. 이는 모델의 네트워크로 하여금 새로운 고해상도의 이미지가 입력될 때, 학습을 위해 해당 이미지의 크기를 조정하기 때문에 필터를 조절할 시간의 확보가 어려움. 이를 해결하기 위해 YOLOv2를 이용하여 필터 조절의 소요 시간을 감소시키고, mAP는 향상시킴

- Convolution with Anchor Boxes는 bounding box의 형식에서 anchor 박스로 변경된 구조로, 기존 YOLO는 Fully Connected layer를 통해 바운딩 박스를 위치와 크기를 예측함. 반면에 YOLOv2는 Faster R-CNN에서 사전에 정의한 다양한 크기 박스의 sliding window 방식임. 이는 각각 class와 region의 예측치를 사용하여 높은 성능의 box를 사용하는 anchor 박스 방식을 사용함
- Dimension Clusters는 기존 anchor 박스를 사용할 때, 박스의 개수를 사용자가 임의로 정해야 하는 문제를 해결하기 위해 k-means 클러스터링을 이용하는 방법임. 만약 k-means의 기본 유클리디언 거리 방식을 사용하면 박스의 크기가 클수록 에러율이 높아지는 경향이 있음

• 에러율 감소를 위해 IoU(Intersection over Union)을 사용한 K-means 거리

$$d\,(anchor, centroid) = 1 - IOU\,(anchor, centroid)$$

- Direct Location Prediction은 anchor 박스 기법을 사용할 때 나타나는 모델의 불안정성을 해결함. 모델의 불안정성은 대부분 box의 x축과 y축을 예측할 때, 가장 많이 나타남. 아래의 식을 통해 anchor 박스의 위치를 예측할 경우, t_x와 t_y의 값이 1씩 변화함에 따라 좌, 우 또는 상, 하로 anchor 박스의 크기만큼 이동함. 이는 anchor 박스의 위치가 이미지 내 어디에든 존재할 수 있게 되므로 랜덤한 값으로 해당 값을 초기화할 경우 offset을 안정화하는데 시간을 소요함. 해당 문제를 해결하기 위해 t의 값의 범위를 제한해야 함

• IOU를 이용한 K-means 거리

$$x = (t_x \times width_a) - x_a$$
$$y = (t_y \times height_a) - y_a$$

- YOLOv2는 로지스틱 활성화 함수를 사용하여 x축, y축, box의 높이, 넓이, 검출된 object의 확률과 IoU를 tx, ty, tw, th, to를 사용하여 0과 1 사이의 값으로 출력되도록 함. 또한 1이상의 anchor 박스의 x, y의 위치값에 대해서 cx, cy를 사용함. 이에따라 Direct Location Prediction을 통해 정확도 향상이 가능함

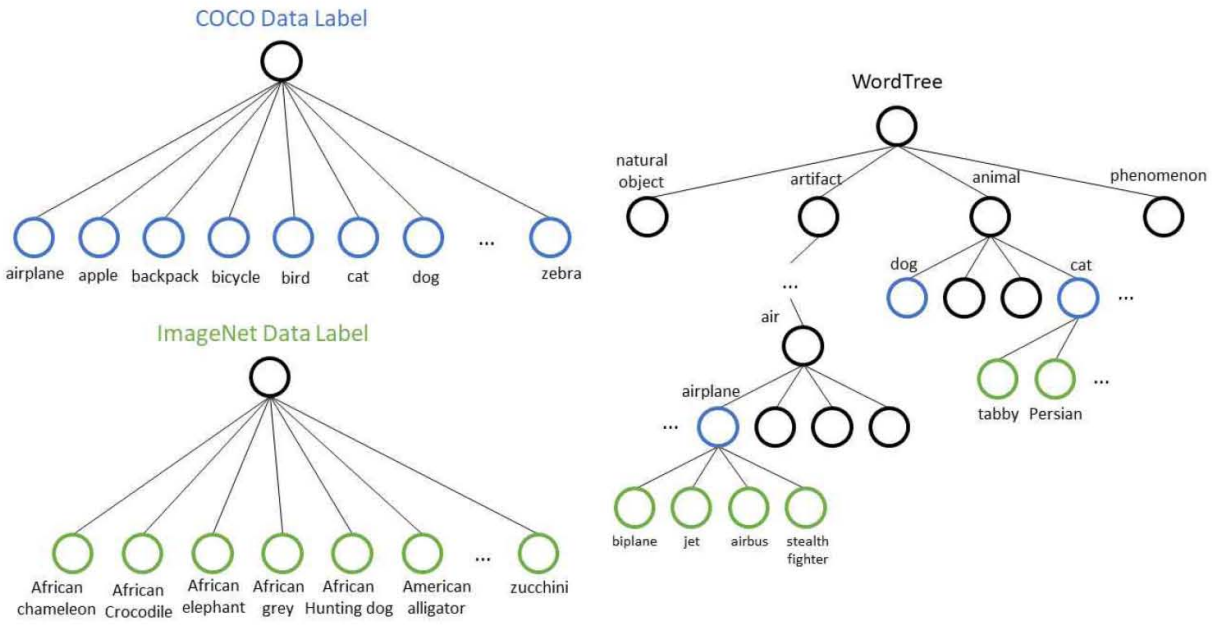
• anchor 박스의 위치를 예측 계산 방법

$$b_x = \sigma(t_x) + C_x$$
$$b_y = \sigma(t_y) + C_y$$
$$b_w = p_w e^{t_w}$$
$$b_h = p_h e^{t_h}$$
$$Pr(Object) \times IOU(b, Object) = \sigma(t_o)$$

- Multi-Scale Training을 통해 다양한 이미지의 크기에 대응하도록 함. 이는 기존의 Fully-Connected layer를 삭제함으로써 네트워크에서 학습 중에 다양한 크기의 이미지를 학습할 수 있도록 함. 이를 통해 고해상도에서 약간의 속도 저하가 일어나지만, 높은 정확도의 객체 탐지가 가능함

- Hierarchical classification은 language database인 WordNet에서 ImageNet 레이블을 가져옴. 언어는 복잡한 구조로 이루어져있기 때문에 directed graph의 형식을 가지고 있음. 각 노드에서 root 노드로 가는 path를 제거하여 tree 구조로 변환할 수 있음. WordTree를 이용하여 Hierchical classification을 진행함. 진행하는 방법은 추출된 객체를 각 자식 노드마다 해당 클래스로 분류될 확률을 계산함
- Dataset Combination with WordTree는 실제 WordTree를 이용하여 COCO 데이터 레이블과 ImageNet 데이터 레이블을 조합함. COCO 데이터 레이블은 일반적인 레이블을 보여주므로, 상위노드에 배치함. ImageNet 데이터 레이블은 하위의 자세한 레이블을 나타내므로 하위노드로 배치함

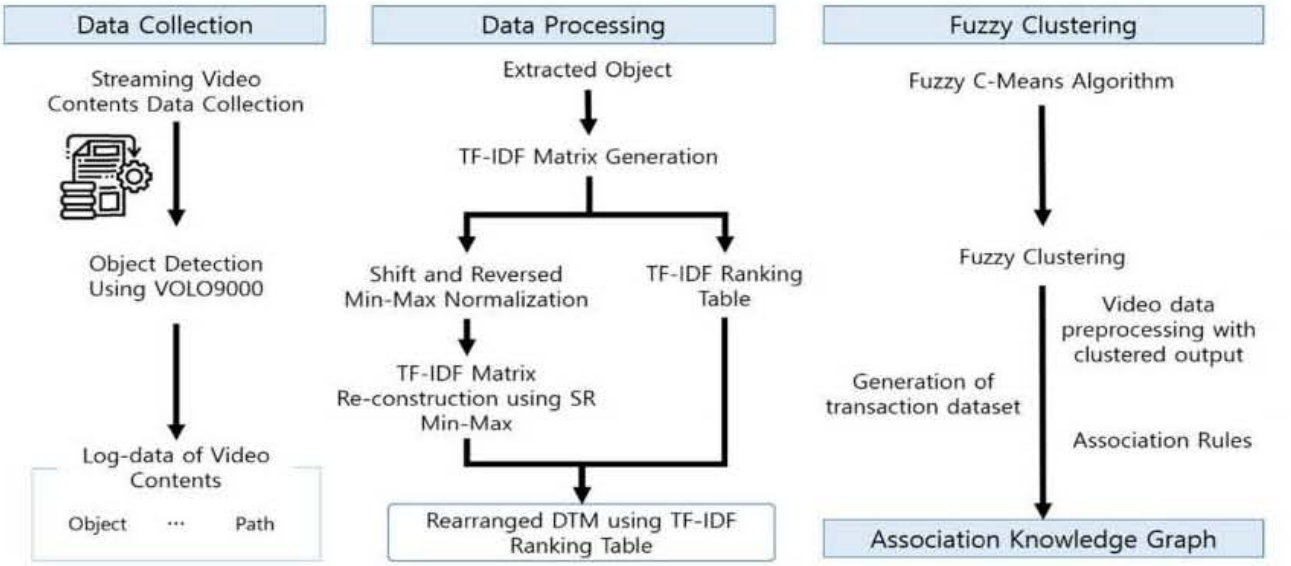
• COCO 데이터 레이블과 ImageNet 데이터 레이블을 WordTree에 조합하는 과정



■ 퍼지 클러스터링 및 최소-최대 정규화를 사용한 연관지식 그래프[45]

- 영상 콘텐츠 데이터는 다양한 분야에서 사용되고, 각기 다른 주제를 제공하지만 숨겨진 유의미한 관계가 존재함. 실시간으로 생성되는 영상 콘텐츠 데이터의 관계를 파악하고, 중요한 규칙 정보를 추출하기 위해 퍼지 클러스터링과 Min-Max 정규화를 이용한 영상 콘텐츠 연관지식 그래프를 개발함

• 퍼지 클러스터링과 Min-Max 정규화를 이용한 영상 콘텐츠 연관지식 그래프 프로세스



- 스트리밍 영상 콘텐츠 데이터를 수집하여 전처리를 진행. 전처리는 YOLO9000을 이용하여 영상 콘텐츠 데이터에서 나타나는 객체를 탐지하여 분류하고, 로그 데이터를 추출
- 추출된 객체를 통해 TF-IDF 매트릭스를 생성 및 영상의 길이를 고려하기 위해 Shift and Reversed Min-Max를 통해 정규화하고, TF-IDF 매트릭스를 재구성하여 Ranking 테이블을 생성

- fuzzy c-means 알고리즘을 이용하여 fuzzy 클러스터링을 진행함. fuzzy 클러스터에서 애매모호한 관계를 추출하기 위한 연관지식 그래프를 생성함. 이를 통해 객체가 다른 영상 콘텐츠에서 애매모호한 관계를 고려함으로써, 연관지식 추출의 정확도를 향상시킴

▪ YOLO 알고리즘을 이용한 객체 정보 수집

- 이머징 리스크의 관계를 파악하기 위해서 670개 영상을 추출함. 추출한 영상은 YOLO9000 알고리즘을 통해 콘텐츠에 존재하는 객체를 분류함. YOLO9000 알고리즘은 객체 검출 속도와 정확도는 YOLOv4 보다 낮지만 한정적 객체 탐지 알고리즘 보다 약 450배 많은 객체를 탐지하여 분류가 가능함[46]
- 추출된 객체는 시간의 흐름에 따라 업데이트되면서 소멸되는 문제가 존재하기 때문에 객체 데이터를 실시간으로 로그 데이터에 기록하는 과정 진행함. 670개 영상 콘텐츠의 영상 스트림의 크기, 입력 영상의 경로, 훈련 시 사용된 이미지의 양, mini batch의 크기, 초당 YOLO9000이 처리한 프레임의 개수, 평균 초당 프레임 개수, 추출된 객체, 추출된 객체의 confidence와 bounding box의 크기 등으로 구성

- 콘텐츠 로그 데이터에서 추출된 객체의 수는 다양하지만 프레임 당 객체들이 존재함. 추출에 정확도가 낮게 분류된 객체들이 존재하기 때문에 Confidence가 50% 미만인 객체를 제거하는 전처리 과정을 진행함. 객체의 이름은 프레임마다 추출된 객체의 집합이므로 중복 가능

45) H. J. Kim, J. W. Baek, K. Chung (2021) Associative Knowledge Graph Using Fuzzy Clustering and Min-Max Normalization in Video Contents, IEEE Access, 9, 74802-74816.
46) 백지원, 김민정, 권혜정, 강지수, 정경용, YOLO를 이용한 객체 지식 그래프 기반 이상 탐지, 2021 한국인공지능융합기술학회 하계학술대회, pp. 27-28, 2021.

• 영상 콘텐츠의 로그 데이터

Variable \ Video	Video_1	Video_2	…	Video_670
Video stream	640x360	640x360	…	640x360
Video file Path	./input/national/0.mp4	./input/national/1.mp4	…	./input/national/669.mp4
Trained mini_batch Size	15,501	15,501	…	32,032
AVG_FPS	1	1	…	1
Objects	3.8	3.8	…	3.8
…	African person combatant trainer attendant worker lifeguard traveler woman …	crab vertebrate animal arthropod decapod crustacean axolotl nutriment edible fruit vegetable …	…	pottedplant apple bird cat person tvmonitor keyboard cake …

■ **Shift and Reversed Min-Max를 이용한 정규화 및 전처리 과정**

- 연관규칙 생성을 위해 사용하는 트랜잭션은 Term Frequency가 고려되지 않는 문제점이 존재하기 때문에 영상 콘텐츠에 등장하는 객체의 출현 횟수를 고려하지 않음. 영상 콘텐츠에 등장하는 객체 출현 횟수를 고려하기 위해 TF-IDF 기반으로 생성한 랭킹 테이블 사용함
- 각각의 영상 콘텐츠에서 중요한 객체의 정보를 유지하고, 불필요한 데이터를 제거하여 최적화된 지식 그래프를 생성함. 이는 추출된 데이터를 통해 각 영상 콘텐츠를 하나의 문서로 분류하여 Document Term Matrix를 생성함. 각 Document Term Matrix를 병합하여 모든 영상 콘텐츠에 대한 Matrix를 생성함

- TF-IDF 가중치를 적용하기 위해 각 영상 콘텐츠에서 등장한 객체의 횟수와 모든 영상 콘텐츠에서 해당 객체의 등장 횟수를 계산하여 Term Frequency×Inverse Document Frequency를 계산 결과를 통해 Matrix의 가중치를 갱신

• 추출한 객체를 이용한 TF-IDF 기반 Document Term Matrix

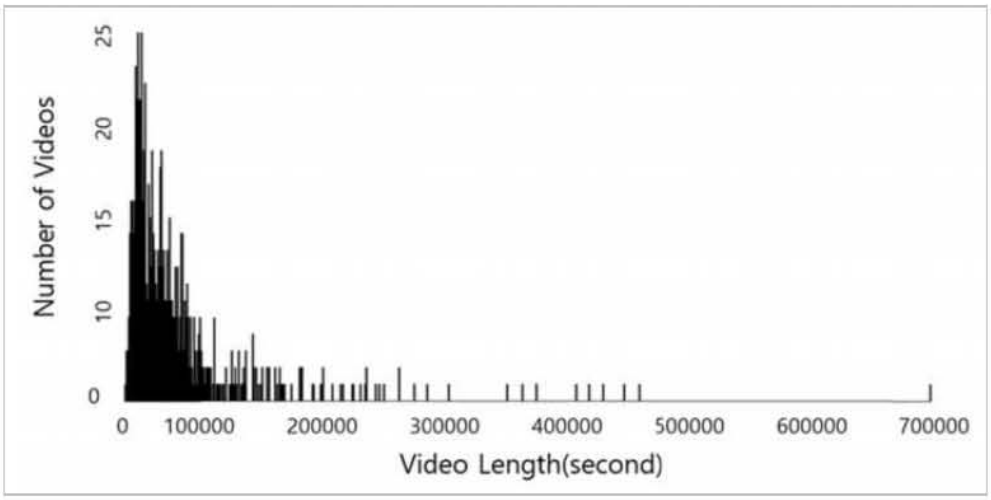
Video \ Object	african	artifact	attendant	bird	…
Video_1	0.002533	0.001210	0.010434	0.003530	…
Video_2	0.004726	0.004513	0	0.054332	…
Video_3	0	0	0.012675	0	…
Video_4	0	0.110116	0	0	…
Video_5	0	0	0	0	…
…	…	…	…	…	…

- 문서에 따라 각 Term의 가중치를 모두 합하여 TF-IDF Ranking Table을 생성 가능함. 객체 추출 알고리즘은 각 프레임마다 객체를 추출하기 때문에 영상 콘텐츠의 길이가 길수록 프레임 수가 증가하기 때문에 특정 객체가 빈번하게 추출될 확률 증가함
- 상대적으로 길이가 짧은 영상 콘텐츠는 중요한 객체를 추출하지만 적게 등장하기 때문에 길이가 긴 영상 콘텐츠에 중요하지 않은 객체의 등장 횟수가 길이가 짧은 영상 콘텐츠보다 더 많이 등장하는 문제 존재
- 이머징 리스크에 대한 발생 가능한 유의미한 관계 추출을 위해 영상 콘텐츠의 길이 데이터를 이용하여 reversed min-max 정규화 진행

- 길이가 가장 긴 영상의 경우 0의 값으로 결정이 되어 추출된 객체들이 완전히 무시될 수 있기 때문에 양수로 1만큼 시프트하여 짧은 영상에게 더 많은 가중치를 주고 긴 영상에게는 가중치를 부여하지 않는 정규화를 진행

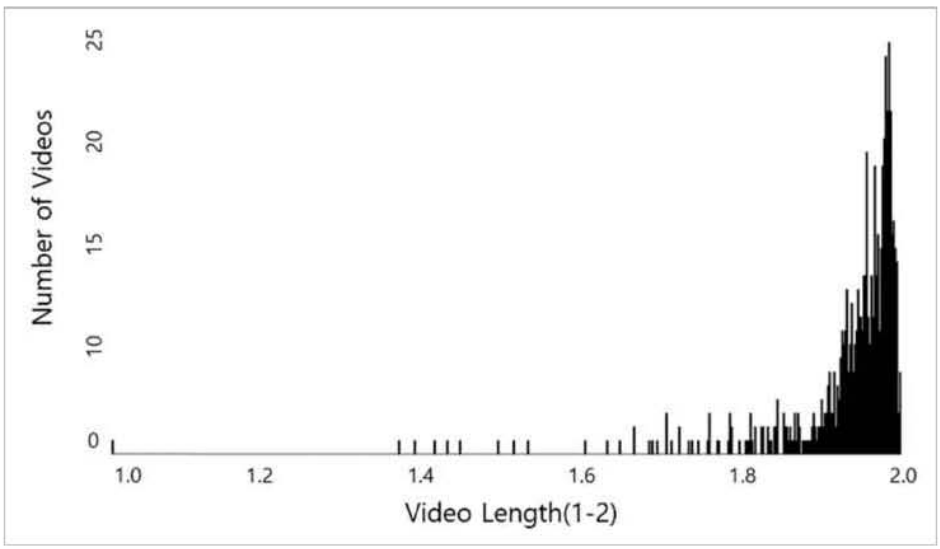
- Min-Max Normalization를 이용하여 전체 데이터 집합 x에서 모든 원소의 값을 0과 1사이로 나타냄. 최댓값과 최솟값의 차를 분모로 하여 데이터의 범위를 결정하고, 분자에 각 x의 원소에 x 원소 중 최소값과 빼는 작업을 진행하면 각 원소를 0과 1 사이로 나타내는 것이 가능함
- Shift and Reversed Min-Max Normalization은 분자를 x의 원소 중 최댓값에 각 x원소를 빼어 큰 값을 1에 가깝게, 작은 값을 2에 가깝게 범위를 지정하는 것이 가능함

• 영상 콘텐츠의 길이와 각 길이에 해당되는 영상 콘텐츠의 개수



- 수집한 콘텐츠가 0에서 10,000초 사이의 영상 콘텐츠의 개수가 10,000 ~ 70,000초의 길이를 가진 개수보다 많은 것을 알 수 있음. 영상 콘텐츠의 길이가 긴, 즉 프레임의 개수가 많은 소수의 영상이 다량의 짧은 영상에서 중요한 객체의 정보를 추출할 수 있는 문제가 존재하기 때문에 Shift and Reversed Min-Max Normalization 진행함

• Shift and Reversed Min-Max Normalization을 적용한 영상 길이 데이터



- 1과 2 사이로 정규화가 진행된 그래프는 x축의 범위가 0에서 70,000초의 범위가 1에서 2로 변환되었기 때문에 영상의 개수를 나타내는 y축의 범위가 22에서 30으로 소폭 상승함
- 길이가 70,000초로 가장 길었던 영상이 가장 낮은 가중치인 1.0을 부여받고, 짧은 영상이 1.8에서 2.0의 높은 가중치를 부여받음으로써 영상의 길이에 따라 다른 가중치를 부여하는 것이 가능하기 때문에 각 영상

콘텐츠가 부여받은 Shift and Reversed Min-Max Normalization 가중치를 TF-IDF 가중치가 계산된 Document Term Matrix에 적용

• Shift and Reversed Min-Max Normalization을 반영한 최종 Document Term Matrix

Object / Video	african	artifact	attendant	bird	…
Video_1	0.005004	0.002390	0.020613	0.006975	…
Video_2	0.009348	0.008928	0	0.107469	…
Video_3	0	0	0.25092	0	…
Video_4	0	0.217403	0	0	…
Video_5	0	0	0	0	…
…	…	…	…	…	…

- Document Term Matrix의 각 단어의 순위를 평가하기 위해 해당 객체의 가중치 값을 전부 합하여 단어의 중요도를 산정하여 가장 높은 값부터 내림차순을 진행하여 최종적으로 TF-IDF 기반 순위 테이블을 생성

• TF-IDF 기반 순위 테이블

Rank	Object	Score
1	africanelephant	40.006615
2	car	37.933949
3	person	34.853192
4	display	20.685892
5	icebear	19.845954
…	…	…

■ 연관지식 그래프 추출을 위한 퍼지 군집

- Shift and Reversed Min-Max Normalization을 통해 재구성한 Document Term Matrix에는 각 영상에서 객체의 중요도에 따른 가중치 정보가 존재함. 영상 콘텐츠의 주제에 따라 동물의 자연 및 생활환경과 인간의 문화에서 연관되어 발생할 수 있는 이머징 리스크의 관계를 추출하기 위한 지식 그래프를 생성
- 이머징 리스크에 대해 애매모호한 관계들이 존재함. Fuzzy C-Means Clustering을 지식 그래프에 적용하여 애매모호한 이머징 리스크의 관계를 파악함. Fuzzy C-Means Clustering은 0 또는 1로만 표현하여 정확히 구분하는 Hard Clustering 기법인 K-means Clustering과 차이가 존재함. Fuzzy C-means Clustering은 모호한 대상을 각 클러스터에 속할 확률을 0~1 사이의 값을 이용하는 Soft Clustering 기법임

• TF-IDF Ranking Table을 통해 재구성한 DTM

TF-IDF Rank	1	2	…	3301	3302
Object	African elephant	car	…	palisade	perch
Video_1	2.96898	0	…	0	0
Video_2	2.75519	0	…	0	0
Video_3	2.66425	0.0178	…	0	0
Video_4	2.32706	0	…	0	0
Video_5	1.88796	0	…	0	0
…	…	…	…	…	…

- 670개의 영상정보에서 중복되는 객체가 등장하지만, 영상의 객체 등장의 개수 등의 차이를 구분하여 이를 기준으로 군집이 가능하고, Fuzzy C-Means Clustering을 위해 Document Term Matrix에서 총 점수가 높은

객체들을 이용함. 추출된 약 3,302개의 객체 중에서 출현 빈도가 낮은 객체의 경우, 해당 객체가 등장하는 영상 콘텐츠의 DTM 행을 제외한 나머지 영상의 행이 0으로 도출

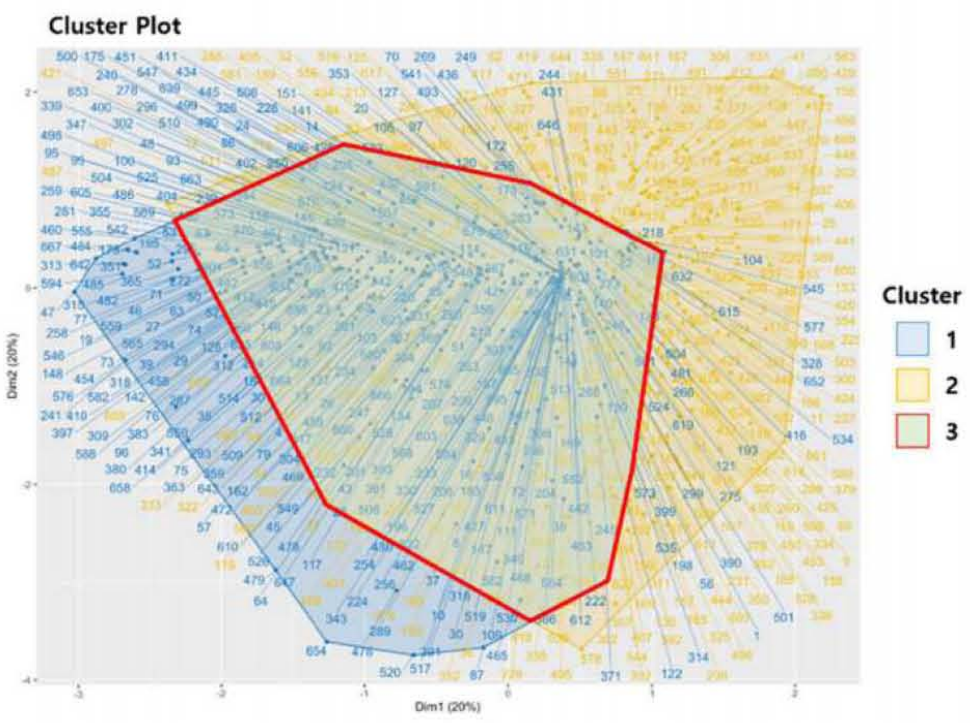
- 무작위 클러스터링 되는 문제가 발생하기 때문에 모든 영상 콘텐츠에서 공통적으로 등장한 객체를 계산하여 3,302개의 객체 중에서 TF-IDF Ranking Table의 순위를 통해 Document Term Matrix를 재구성
- 상위 TF-IDF Ranking을 가진 객체의 경우 하위 TF-IDF Ranking 객체에 비해 가중치 값이 많다는 것을 알 수 있음. 상위 2%의 TF-IDF Ranking의 객체를 이용하여 DTM을 구성하여 Fuzzy C-Means 군집을 수행함. 유클리디언 제곱의 식을 기반으로 클러스터의 중심과 영상 콘텐츠 간의 거리를 계산하여 군집을 진행

• Fuzzy C-Means 군집을 위한 제곱 유클리디안 거리

$$SED = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

- Clustering Center의 개수는 Total within sum of square를 이용한 elbow method와 Silhouette 계수를 이용하여 결정함. Cluster Center가 2일 때, 가장 넓은 Silhouette width를 가지므로 최종적으로 Cluster Center의 개수를 2로 설정함. 이후 총 670개의 영상 데이터를 클러스터링한 결과를 시각화[47]

• Fuzzy C-Means 군집 결과 시각화



- Fuzzy C-Means 군집을 시각화한 그래프에서는 총 두 개의 클러스터가 생성됨. 각 영상 콘텐츠에 대한 번호 라벨과 클러스터의 색이 같은 경우, 클러스터에 해당 영상 콘텐츠가 포함됨

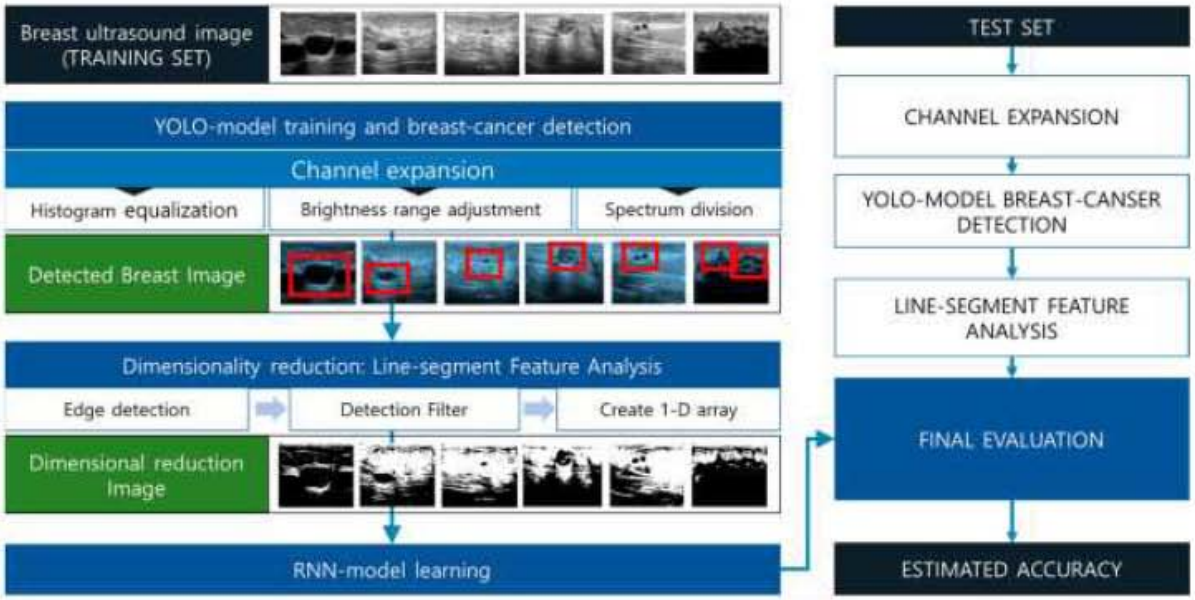
47) H. J. Kim, J. W. Baek, K. Chung (2020) Optimization of Associative Knowledge Graph using TF-IDF based Ranking Score, Applied Sciences, 10(13), 4590-4610.

■ 다중 모델 기반 이미지 채널 확장 및 시각적 패턴 표준화를 사용한 건강 위험 감지 및 분류 모델[48]

▪ 다중 모델 기반 초음파 영상 채널 확장 알고리즘 개발

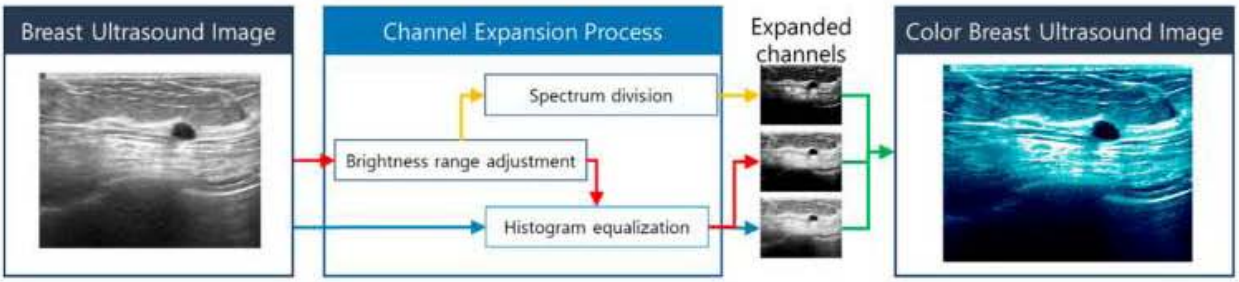
- 반사 및 투사된 신호를 밝기 값으로 변환하여 생성된 데이터이며 밝기에 대해 강한 노이즈를 보임. 의료 영상은 상대적으로 높은 해상도를 나타내기 때문에 학습 모델을 통한 진단 과정에는 고성능 하드웨어가 요구됨. 밝기 노이즈는 탐지 성능 저하를 유발할 수 있으며, 고해상도는 많은 연산을 한 번에 고속으로 처리해야 하는 고계산 문제를 유발

• 다중 모델 기반 이미지 시각적 패턴 표준화 및 차원 축소 알고리즘의 프로세스



- 다중 모델 기반 영상 채널 확장 및 시각 패턴 표준화 알고리즘을 사용함. 채널 확장을 통해 초음파 영상의 경계 영역을 강화하고 약한 노이즈 제거 기법을 사용하여 초음파 영상을 더 선명하게 처리함

• 이미지 경계를 강화하기 위한 채널 확장 프로세스



- 확장된 데이터를 YOLO 모델에 적용하여 검출 성능을 향상함. 검출된 객체 영상을 시각 정보를 이용한 정규화를 통해 축소 데이터로 처리하여 RNN 모델에 대한 입력으로 사용함
- 이미지 채널 확장 및 전처리를 수행하여 영상의 노이즈를 효과적으로 제거하고 내부 경계를 강화하는 과정을 통해 YOLO의 검출 성능을 향상함. 채널 확장은 히스토그램 정규화, 밝기 범위 조정을 통한 밝기 균등화, 임계값을 통한 밝기 영역 제어를 통해 수행되어 영상의 밝기 영역 간 간격을 넓힘
- 임계값을 0~255 범위에서의 임계값 스펙트럼 영역을 사용하여 각 레벨의 밝기 범위로 N(N=5)개의 레벨로 나누어서 일반적으로 단위형으로 구성된 영상에서 최대 밝기를 255로 계산함. 입력에 분포된 밝기 값은 계산된 스펙트럼 영역을 사용하여 비교함

48) C. M. Kim, E. J. Hong, K. Chung, R. C. Park (2021) Health Risk Detection and Classification Model Using Multi-Model-Based Image Channel Expansion and Visual Pattern Standardization, Applied Sciences, 11(18), 8621-8640.

• 밝기 스펙트럼 분할 알고리즘

Algorithm 1 Brightness Spectrum Division Algorithm

```
Input: x
def Extraction of Feature:
  LABEL = [0,2,4,8,16]
  SPECTRUMAREA = 255//N
  for i from 0 to LEN(LABEL) do
    S = copy(x)
    S[SPECTRUMAREA xi > S] = 0
    S[SPECTRUMAREA x(i+1) < S] = 0
    S[S != 0] = LABEL[i]
    S = ((S/MAX(LABEL)x255)
    Y += S
Output: Y
```

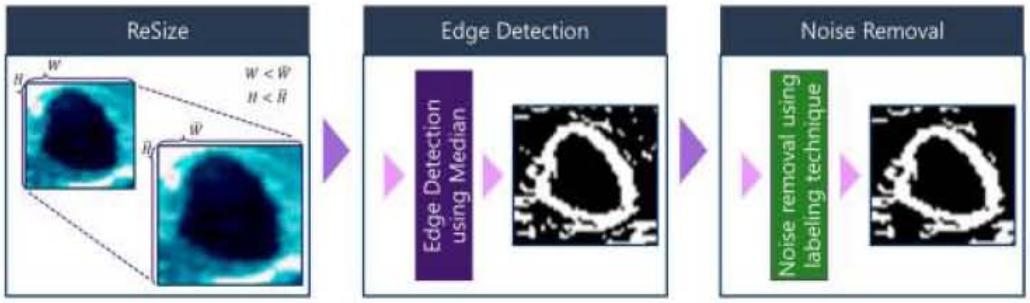
- 확장된 채널이 있는 이미지는 원본 이미지보다 더 강한 경계를 보여주며, 약한 밝기 영역과 강한 밝기 영역이 명확하게 구분됨. 확장된 채널 이미지는 YOLO 모델의 학습 데이터로 사용되며 영상에서 객체 검출에 사용함

■ 차원 축소를 사용한 시각적 패턴 표준화

- 원본 영상보다 더 명확한 경계를 가진 확장된 채널 영상을 생성하더라도 리소스 사용 제한으로 인해 정확한 분석이 어렵거나 아예 불가능한 경우의 발생 가능성이 존재함. Yolo 모델의 탐지성능을 향상하기 위해서 이미지 차원 축소와 RNN 모델 학습을 진행함

- RNN 기반 시각 정보를 이용한 정규화된 패턴 변환을 통한 진단 모델을 사용함. YOLO 모델을 통해 위치를 감지하고 해당 위치를 관심 영역(ROI)으로 설정하여 추출함. 사전 정의된 감지 필터를 사용하여 이미지의 시각적 정보를 일련의 정규화된 패턴으로 변환함. 이는 정규화된 데이터에서 유형을 분류하기 위한 RNN 모델의 입력으로 사용되는 N×M 행렬로 정의됨
- YOLO를 통해 식별된 영상의 윤곽을 감지함. 윈도우를 이용해 중앙값을 통한 윤곽 검출을 진행함. 검출된 영상의 크기를 변경한 후 윤곽 검출 및 노이즈 제거를 수행함. 이미지 크기를 $N \times N$으로 분할 할 때 이미지 크기를 16등분 하도록 조정함

• 이미지 크기 조정 및 가장자리 감지 결과



- 검출된 영역만을 ROI에 의한 노이즈 제거를 통해, 이진 윤곽 검출을 수행하고 계산된 이진 윤곽 영상은 차원 축소를 위해 특정 형태를 표현하기 위해 필요한 선분의 종류와 개수로 일련의 수치 패턴으로 정규화하며 동일한 패턴을 나타내는 정보를 누적하여 집계를 진행

- 내부 영역의 노이즈 제거를 위해 전처리된 이미지의 ROI 없이 원본 이미지의 윤곽 감지 과정에서 이미지를 생성함. 윤곽 검출 및 이진화 과정을 통해 계산된 이미지에 나타나는 작은 점과 같은 노이즈를 레이블링 알고리즘을 통해 내부의 노이즈를 제거함

• 레이블링 알고리즘

Algorithm 2 Labeling Algorithm

```
Input: X
def Labeling:
    A = X.copy()
    H, W = X.shape[]
    pPut, labelNum, LABEL, CHECK = [], {},0, False
    for h from 0 to H do
      for w from 0 to W do
        if A[h, w] is 255:
          pPut.put((h, w))
          while LEN(pPut) > 0 do
            CHECK = True
            n, m = pPut.pop()
            A[n, m] = 0
            if LABEL in labelNum:
              labelNum[LABEL] = labelNum[LABEL].put((n, m))
            for i from 0 to 3 do
              for j from 0 to 3 do
                if A[n+i, m+j] is 255:
                  pPut.put((n+i, m+j))
        if CHECK:
          CHECK = False
          LABEL += 1
Output: labelNum
```

- 시각적 정보를 일련의 패턴으로 변환하기 위해 영상 조각으로 분할을 진행함. 최종 결과를 계산하기 위해 세부 데이터를 정렬하는 기준 순서로 지정됨. 외부 형태를 인식하여 내부 물체를 인식한다는 사실을 활용하여 이미지 시퀀스를 이미지의 가장 바깥쪽에서 안쪽 영역으로 배열함
- 탐지 필터를 사용하여 정규화를 위해 시각적 정보를 일련의 숫자 패턴으로 변환하며 이미지의 실제 픽셀 수에 해당하는 숫자 패턴을 얻음

3. 연구개발과제의 수행 결과 및 목표 달성 정도

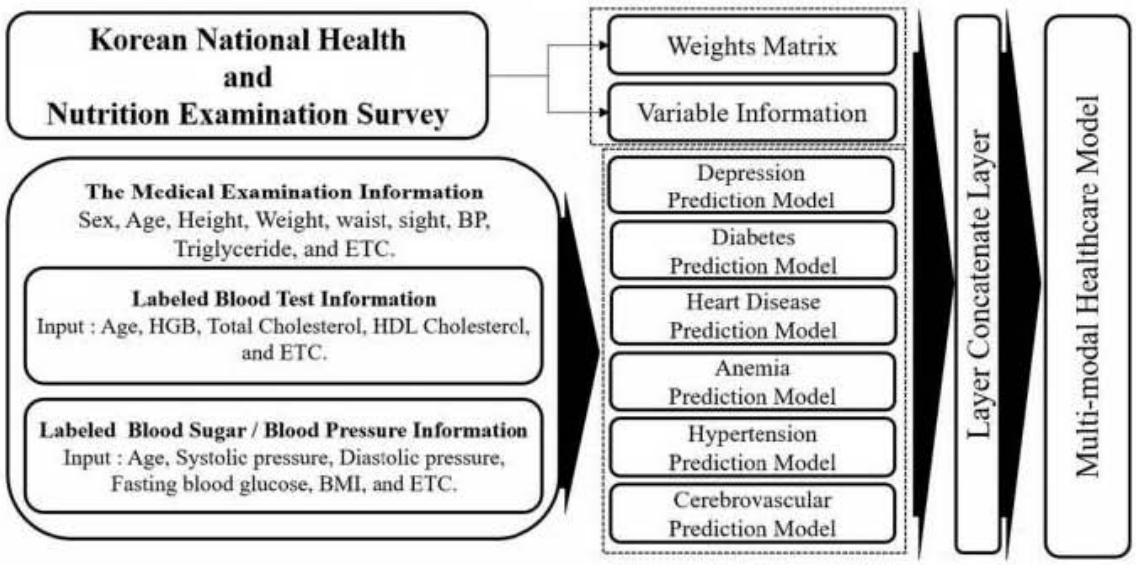
1) 연구수행 결과

(1) 정성적 연구개발성과

■ Collaborative Concat Layer를 이용한 하이브리드 멀티-모달 딥러닝 성능평가[49]

- Collaborative Concat Layer(CCL)은 머신러닝 기법을 통해 변수의 특징을 분석하고, 분석결과 나타나는 관계를 이용하는 방법을 개발하였고, 평가하기 위한 실험으로 사전에 학습된 싱글-모달 헬스 모델을 CCL로 연결하여 멀티-모달 모델을 생성하고 정확도 평가를 진행함
- 구성된 싱글-모달 모델을 CCL로 연결하였을 때, 기존 모델으로부터의 성능 유지 정도에 대한 성능평가를 진행함. 성능평가는 헬스 데이터를 학습한 DNN 질병 예측 모델에 제안 방법을 적용했을 때, Accuracy 척도 평가를 진행하여 기존 대비 10% 이상 성능평가 결과를 도출

▪ 하이브리드 멀티-모달 헬스 딥러닝에 대한 성능평가를 위한 모델 구성



- 성능평가를 위한 데이터는 국민건강영양조사, 건강검진정보, 국가건강검진 혈압혈당데이터, 국가건강검진 혈액검사 데이터를 사용함. 이 데이터는 디바이스 사용 여부등의 상황적요인에 따라 수집되는 변수들로 구성됨. 국민건강보험 가입자들이 일정 주기로 시행되는 검진정보를 활용함
- 이머징 헬스 리스크에 대한 판단을 위한 환자의 신체적 정보 및 수치를 활용함. 국가건강검진 혈압혈당데이터는 당뇨 및 고혈압의 유병 요인 및 상관변수로 이루어진 데이터임. 국가건강검진 혈액검사 데이터는 혈액 검사 데이터 및 뇌혈관질환, 빈혈 유병여부와 상관된 변수를 포함
- 총 백 만개의 행으로 이루어진 혈액 검사 데이터를 분할하여 60만개는 예측 모델의 훈련단계에서 사용하며 그의 5분의 1인 12만개는 모델 검증에 사용함. 나머지는 테스트 데이터 집합으로 사용함. 이를 통해 3가지의 DNN 예측 모델(심장질환, 뇌혈관질환, 빈혈)을 생성함. 세 모델의 입력 데이터로 사용되는 환자의 신체적 정보 및 혈액 검사 수치는 모두 동일한 변수를 사용함
- 심장질환 예측 모델은 Input(6) × dense(64, relu) × dense(64, relu) × dense(64, relu) × Output(1,Sigmoid)로 구성됨. epoch는 10, 정확도는 0.725로 나타남
- Cerebrovascular 예측 모델은 Input(6) × dense(4, relu) × dense(6, relu) × dense(8, relu) × Output(4, sigmoid)로 구성됨. epoch 5일 때 모델의 정확도는 0.733으로 나타남

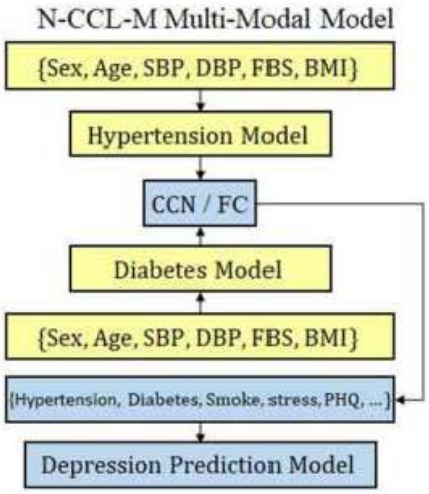
49) J. C. Kim, K. Chung (2020) Hybrid Multi-modal Deep Learning using Collaborative Concat Layer in Health Bigdata, IEEE Access, 8, 192469-192480.

- anemia 예측 모델은 Input(6) × dense(16, relu) × dense(16, relu) × dense(16, relu) × Output(1, Sigmoid)로 구성됨. epoch는 20, 정확도는 0.762로 나타남
- 혈압 혈당 데이터는 100만 개의 Row로 구성되어 있으며, 이 중 60만개를 모델 학습에 사용하고, 그중 20%인 약 12만 개를 validation 데이터로 사용함. 이를 통해 2가지의 DNN 예측 모델(당뇨, 고혈압)을 생성함. 이에 혈압 및 혈당 검사 수치와 수치 변화에 유의미한 관계가 있는 변수(성별, 나이, BMI)를 입력으로 사용해 두가지 모델에 동일하게 적용함
- 당뇨 예측 모델은 Input(6) × dense(64, relu) × dense(64, relu) × Output(1, Sigmoid)으로 구성됨. epoch는 5, 정확도는 0.960으로 나타나고, 고혈압 예측 모델은 Input(6) × dense(64, relu) × dense(64, relu) × dense(64, relu) × Output(1, Sigmoid)으로 구성됨. epoch는 10, 정확도는 0.971로 나타남
- 국민건강영양조사 데이터는 정제와 가공을 통해 선별한 10,488건을 사용하여, 임의로 선별된 전체의 70%에 해당하는 데이터를 분석 및 학습에 사용하고 나머지 3,146건을 평가에 사용함으로써 우울증 예측 모델을 생성함. 우울증 예측 모델은 유병여부, 흡연 여부, 스트레스 검사 변수를 입력으로 우울증을 예측하는 DNN 모델이며, Input(14) × dense(64, relu) × dense(64, relu) × dense(64, relu) × Output(1, Sigmoid)으로 구성됨. epoch는 20, 정확도는 0.956으로 나타남

▪ CCL로 Concatenate하는 방법과 Fully Connected로 Concatenate하는 방법의 정확도 평가[50]

- DNN 모델을 CCL과 FC를 이용하여 N-CCL-M, N-CCL, CCL-M 형태로 Concatenate하여 생성한 멀티-모달 헬스 모델의 성능평가를 진행함
- 실험 1은 입력변수 n개가 부재중인 상황을 가정하고, N-CCL-M 형태로 모델을 연결했을 때 정확도 변화를 평가하였고, 실험 2는 입력변수 n개가 부재중인 상황을 가정하고 N-CCL, 형태로 모델을 연결했을 때 정확도 변화를 평가함. 실험 3은 CCL-M 방식을 사용함. 이는 모델 연결 시 새로이 나타나는 변수에 대한 Accuracy 척도를 평가함. 실험데이터는 국민건강영양조사 데이터에서 평가데이터로 나눈 3,146건을 사용하함. Fully Connected 방식의 Concatenation은 학습 과정이 필요하므로 분석데이터로 사용되었던 7,342건으로 별도 학습을 진행함. 각 실험을 10회 반복하여 평균 정확도를 평가함

• N-CCL-M 멀티-모달 헬스 모델



- 고혈압 예측 모델, 당뇨 예측 모델의 출력 및 PHQ, 스트레스 인자를 통해 우울증을 예측하는 N-CCL-M 방법 기반의 멀티-모달 헬스 모델임. 첫 번째 실험의 경우 환자가 스트레스 정도를 인지하고 있는 정도 및 흡연 여부를 결측되었다고 가정함. 협력적 레이어 혹은 Fully Connected layer를 통해 결측값 예측을 진행하며 이에 대한 입력값은 심장질환, 당뇨병 예측 모델의 출력임. 결측값 예측의 진행 여부에 따른 결과 정확도 비교가 진행됨

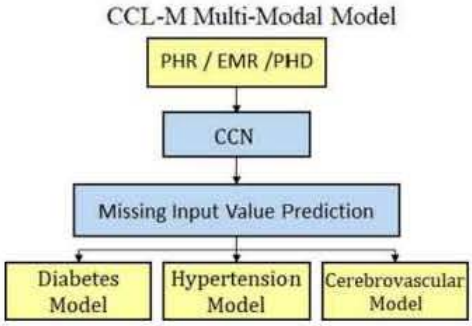
50) 김주창, 예측 정확도 향상을 위한 협력적 레이어 기반 하이브리드 멀티-모달 딥러닝, 경기대학교 컴퓨터과학과, 박사학위논문, 2020.12.

• 성능평가 실험 1의 결과

Models	Accuracy			
Variables	All	-stress	-smoke	-stress, -smoke
Diabetes	0.956	0.656	0.711	0.574
MM N-CCL-M	0.933	0.874	0.891	0.859
MM FC0	0.651	0.434	0.472	0.394
MM FC30	0.860	0.518	0.536	0.579
MM FC50	0.942	0.665	0.681	0.596

- 모든 변수가 가용 상태임을 확인하고, -stress, -smoke는 각 변수를 결측값으로 처리했음을 나타냄. 각 모델의 이름은 멀티-모달 모델(MM), Epoch 횟수(FC n회)로 나타냄
- 기존 모델의 성능을 89% 이상 유지하였으며 각 변수의 결측값 가정에도 불구하고 안정적으로 당뇨 유병을 예측함. FC 사용 모델의 경우 Epoch 횟수와 관련하여 기존 모델과의 유사성 확인이 가능함. 그러나 결측값 가정의 경우 성능 저하가 크게 나타남. 멀티-모달 사용으로 싱글-모달에 비해 높은 정확성을 보이며, 이는 부재하지 않는 다른 변수를 활용한 학습이 가능하기 때문임. 따라서 노드 수 및 초기 가중치 등에 의한 최적화 요구와 데이터가 필요한 한계가 존재하는 FC 기반 모델과 달리 CCL 기반 모델을 통해 유의미한 성능 유지가 가능함을 나타냄

• CCL-M 멀티-모달 헬스 모델



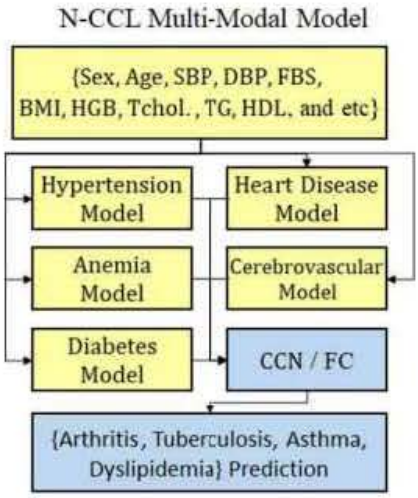
- 상황정보를 기반으로 후위 모델에서 발생하는 결측값 대치를 가능하게 하는 CCL-M 기반 멀티-모달 모델
- 두 번째 실험은 당뇨, 고혈압, 뇌혈관질환 모델 총 3가지의 입력값이 결측되었을 때, 기타 상황정보로부터 산출해낸 예측값을 통한 모델의 정확성을 평가

• 성능평가 실험 2의 결과

Models	Accuracy					
	Diabetes		Hypertension		Cerebrovascular	
Variables	All	-FBS	All	-BP	All	-BP
Single-modal	0.960	0.711	0.971	0.680	0.733	0.571
MM N-CCL-M	0.960	0.918	0.971	0.908	0.733	0.674

- Variables에서 -FBS 및 -BP는 각각 식전혈당 수치와 혈압측정 수치가 결측된 데이터 사용을 의미함
- 성능평가 실험 3에서 싱글-모달 모델은 실험 1과 마찬가지로 입력변수가 결측되면 성능 하락이 크게 나타남. 결측값을 상황정보를 통해 유의미한 관련 값으로 대치함으로써 91%의 정확도를 보장하며, CCL 기반 예측에서 기존 모델의 성능 유지가 가능함을 보임. 모델 복잡도가 증가함에 따라 요구되는 리소스가 많아지는 반면, 결측값이 발생하여 모델의 성능이 하락되는 예외상황을 방지할 수 있음

• N-CCL 멀티-모달 헬스 모델



- 5가지 유병 예측 모델의 출력을 기반으로 전혀 다른 질병 예측 모델(관절염, 결핵, 천식, 이상지질혈증)을 산출하는 N-CCL 기반 멀티-모달 헬스 모델
- 세 번째 성능평가는 N개 모델의 출력을 공통으로 사용하며 CCL 기반의 예측 유무에 따른 멀티-모달 정확성 평가를 진행함

• 성능평가 실험 3의 결과

Models	Accuracy				
Variables	Arthritis	Tuberculosis	Asthma	Dyslipidemia	Avg.
MM N-CCL-M	0.914	0.894	0.876	0.921	0.901
MM FC0	0.684	0.657	0.778	0.818	0.734
MM FC30	0.860	0.889	0.852	0.892	0.873
MM FC50	0.936	0.911	0.878	0.957	0.920

- FC 기반 예측 방법이 CCL 기반의 예측에 비해 평균적으로 0.02의 성능 향상을 나타냄. 그러나 FC 기반 예측 방법 사용의 경우 모델의 추가영역이 최적화 되기 위한 추가 학습이 요구되며 입력값의 변동에 따라 기존 학습 모델을 적용하기 어려운 한계 존재함. 반면 CCL 기반의 예측 방법의 경우 입력값 형태의 변동에도 비교적 간단한 구현을 통해 모델 수정이 이루어지며 추가적인 학습 과정이 요구되지 않아 리소스 절약 면에서 장점 존재함

- 실험 결과 수시로 변화하는 사용자 중심의 헬스 모델에서 CCL을 이용하는 방법이 더 간단하게 모델간의 Concatenation 가능함을 나타냄. 또한, 성능평가 결과 FC 기반 예측 방법의 성능을 재현할 수 있으며 오히려 간단한 구현 및 적은 리소스 사용으로 유의미한 결과를 이끌어낼 수 있음을 나타냄. CCL을 통해 다양한 헬스 예측 모델에 대해 재사용과 Concatenate가 이점이 있음을 나타냄으로써 기존의 딥러닝 모델을 머신러닝 기법으로 연결하는 하이브리드 멀티-모달 딥러닝을 통해 실제 사용자 상황이나 주변 상황에 적합하게 모델을 개선할 수 있음
- CCL은 매우 높은 정확도를 나타내지는 않지만, 추가 학습이나 작업 없이 바로 사용이 가능한 장점이 있음. 또한, 주요 파라미터 변수간의 상관계수나 평균값에 대한 최적화를 통해 성능을 향상시킬 수 있음. CCL은 parameter인 weight matrix와 variable information의 업데이트를 통해 정확도 향상이 가능하며, 새로운 변수가 수집되어 변수간의 상관계수가 추가될 경우 확장성을 가질 수 있음. 현재의 variable information에 있는 각 변수의 평균값은 보편적으로 수집된 데이터에 의한 평균으로, 이는 실제 사용자에 따라 달라질 수 있음

■ **결측치 추론 방법에 따른 성능평가**

- 실험데이터를 멀티-모달 Denoising 오토인코더, K-nn, SVD, column mean(c.mean) 방법으로 처리하고, 이에 따라 학습데이터로 구성하고, 학습 데이터를 동일한 머신러닝 알고리즘에 적용하여 생성되는 모델을 평가
- 국민건강영양조사 데이터를 전처리하여 구성한 14,688건을 활용하고, 임의의 70:10:20의 데이터세트 분할을 통해 각각 학습과정, 검증과정, 테스트 과정에 사용함

- 학습 데이터는 각각의 모델을 생성하는데 사용하고 검증데이터를 이용하여 평가함. 생성된 각각의 모델을 이용하여 임의의 결측값을 생성한 테스트 데이터를 적용. 테스트 데이터는 무작위로 결측값을 생성하며 노이즈 요인은 0.05와 0.30 범위 내에서 0.05의 간격으로 결정됨. 출력값과 원본 데이터의 오차 측정을 진행하며, 결측치와 원본 데이터간 비교를 통해 오차 정도를 평가함

● 결측치 추론 방법에 따른 성능평가 결과

Missing Values Estimation Method (Accuracy)

Noise factor	SVD	K-NN	c.mean	SDAE	MMSDAE
0.05	0.9142	0.8901	0.7144	0.9675	0.9617
0.10	0.8941	0.8764	0.6751	0.9619	0.9571
0.15	0.8595	0.8490	0.6214	0.9556	0.9459
0.20	0.8012	0.7907	0.5451	0.9458	0.9345
0.25	0.7427	0.7242	0.5048	0.9321	0.9217
0.30	0.7019	0.6726	0.4191	0.8790	0.8610

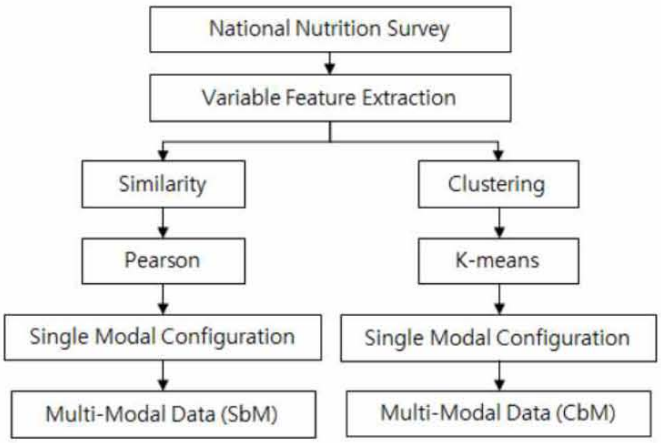
- 노이즈 요인에 따른 모델별 결측치 추론의 정확도를 나타내고, 노이즈 요인과 상관없이 오토인코더를 이용한 결측치 추론이 가장 유의미한 결과를 나타내며, 이는 노이즈 요인이 높을수록 정확도가 낮게 나타남

■ **멀티-모달 구성 방법에 따른 성능평가**

- 머신러닝은 변수 구성이나 학습 데이터에 따라 학습효율, 학습시간 등이 변화하며, 모든 변수를 하나의 알고리즘으로 처리함으로써 모델의 복잡성을 줄여 비교적 간단한 구현과정을 거치는 장점이 존재함
- 알고리즘의 계산량이 높아져 이를 실제 활용할 때의 효율이 떨어지기 때문에 서로 의미가 명확하지 않은 입력변수간의 관계를 학습에 따라 정확도나 연산속도 저하가 발생. 대량의 입력변수를 가지는 데이터 세트는 변수간의 특징을 분석하여 상호 영향력이 높은 변수끼리 구성하는 계획적인 변수 구성이 필요함
- 유사도, 군집분석을 사용하여 변수간의 특징을 분석하며, 각각의 방법을 이용한 변수 조합을 멀티-모달 데이터로 구성하기 때문에, 유사도 기반 멀티-모달 데이터, 군집 기반 멀티-모달 데이터, 범주 기반 멀티-모달 데이터로 총 3개의 멀티-모달 데이터셋으로 구성함

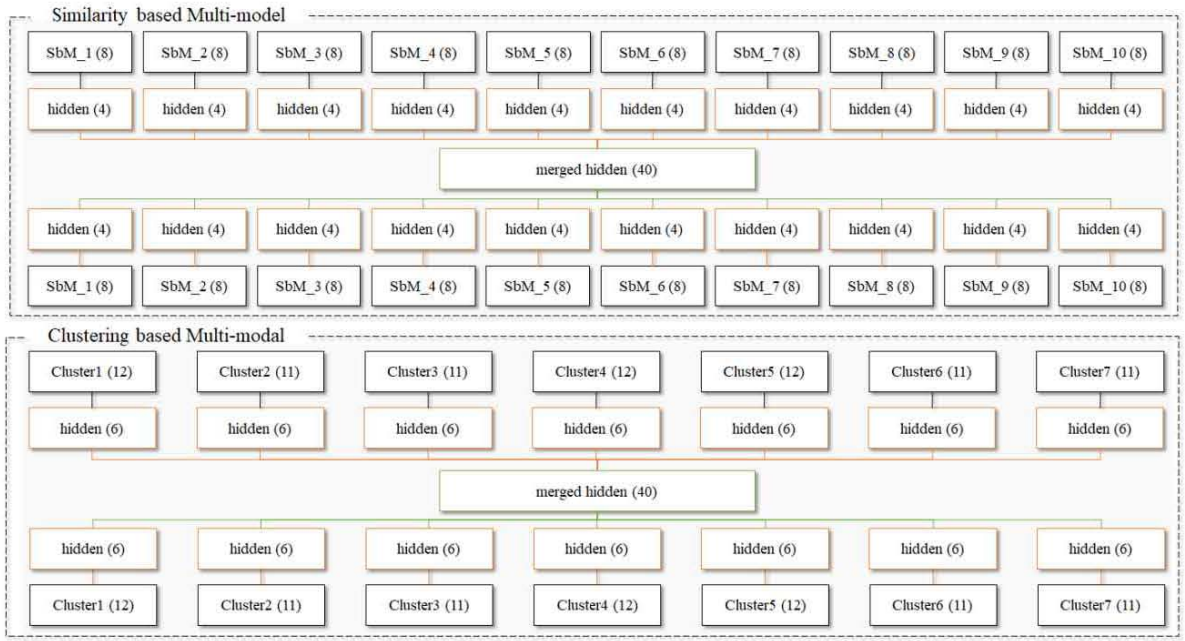
- 유사도 기반 멀티-모달 denoising 오토인코더(SbM_DAE), clustering 기반 멀티-모달 denoising 오토인코더(CbM_DAE), hierachically structure 기반 멀티-모달 denoising 오토인코더(HbM_DAE)에 적합한 모델을 생성하고 성능을 평가함

● 특징 분석 방법에 따른 멀티-모달 데이터 구성



- 유사도 분석은 변수들이 서로 양의 상관관계인지 음의 상관관계인지 판단하는 방법이며, -1부터 +1까지로 나타내며 -1에 가까울수록 서로 반대 방향으로 나아가는 관계이고, +1에 가까울수록 서로 같은 방향으로 나아가는 관계임을 나타내고, 0에 가까울수록 서로에게 미치는 영향력이 작음을 나타냄
- 피어슨 상관계수는 구현이 간단하고 보편적으로 널리 알려진 계수이며, 군집분석은 서로 벡터 공간상에서 거리가 가까운 변수들을 군집으로 구성하고, k-means 알고리즘을 이용하며, 전체 군집의 분산이 가장 작을 때의 k값을 선택함
- 범주 분석은 국민건강영양조사에서 동일한 클래스로 분류된 변수들을 그룹으로 묶는 방법이며, 가장 보편적이고 기본적인 군집 알고리즘으로 구현 난이도에 비해 우수한 성능을 나타냄

● 데이터 구성에 따른 멀티-모달 오토인코더



- 멀티-모달 데이터를 구성하기 위한 다양한 방법이 있으며, 성능평가를 통해 멀티-모달 데이터 구성 방법이 모델의 정확도에 미치는 영향이 있음을 확인하고, 향후 다양한 알고리즘의 적용을 통해 데이터에 따른 알고리즘의 선택이 필요함을 확인함
- 유사도 기반 데이터 구성은 피어슨 상관계수를 사용하며, 랜덤하게 하나의 변수를 선택하고 유사도의 절대값이 높은 8개를 그룹으로 묶으며, 선택된 변수를 제외하고 남은 변수들을 이용하여 이를 반복함으로써, 8개의 변수로 구성된 10개의 그룹으로 구성함
- k-means에서 k값은 편차의 제곱이 가장 작게 나타나는 7을 k값으로 사용하고, 범주 기반 구성은 의사진단 여부(dg), 검진조사(HE)와 같이 데이터의 라벨에 따라 구성하며, 5개의 그룹으로 구성되며, 히든레이어 노드

의 크기는 입력값을 반으로 나눈 값의 반올림한 값으로 구성함
- 레이어 SbM_1(8)은 유사도 기반 멀티-모달 데이터에서 그룹1에 8개의 입력노드가 있음을 나타냄.
- 멀티-모달 오토인코더에서 변수의 구성에 따라 히든 노드의 수를 제외한 조건은 모두 동일한 조건에서 학습하고, 멀티-모달 Denoising 오토인코더는 싱글-모달에서 히든 레이어를 3개로 구성하며, 각각의 싱글-모달을 통합하는 과정이 필요하여 전체 가중치 수를 조절함

● 멀티-모달 denoising 오토인코더 학습 결과

SbM		CbM		HbM	
Parameter	5,032	Parameter	6,488	Parameter	4,790
Noise Factor	Accuracy	Noise Factor	Accuracy	Noise Factor	Accuracy
0.05	0.9601	0.05	0.9577	0.05	0.9617
0.10	0.9559	0.10	0.9510	0.10	0.9571
0.15	0.9412	0.15	0.9426	0.15	0.9459
0.20	0.9313	0.20	0.9312	0.20	0.9345
0.25	0.9209	0.25	0.9241	0.25	0.9217
0.30	0.8497	0.30	0.8550	0.30	0.8610

- 파라미터는 오토인코더의 입력, 히든, 출력 노드의 구성에 따라 결정되며, HbM이 가장 작게 나타나며, 모델의 실제 적용에서 가작 적은 자원을 요구함을 나타냄

- 노이즈 요인에 따른 정확도 역시 HbM이 가장 우수하게 나타나며, 국민건강영양조사에서 구분한 Class가 실제 상황을 고려하여 구성되어 있어 다른 모델에 비해 높은 정확도가 나타나는 것으로 예상됨. Class가 명확하게 구분되지 않은 데이터 세트를 멀티-모달로 구성할 경우 다른 결과가 나타날 수 있음

■ 향상된 AnoGAN을 통한 이머징 리스크에 대한 이상치 탐지에 대한 성능평가[51]

- 이머징 리스크 상황에 따른 심전도 실험을 위한 재분류한 Normal(N), Supraventricular(S), Ventricular(V), Fusion(F), Unknown(Q)로 기준으로 데이터를 구성하고, 전처리하고 모델링하여 성능평가 진행

● 학습데이터와 검증데이터의 구성

AAMI Class	Preprocess Data	Train Data	Test Data
Normal (N)	86717	85717	1000
Supraventricular(S)	3026	-	330
Ventricular(V)	7008	-	330
Fusion(F)	802	-	330
Unknown(Q)	15	-	15

- 모델링 단계에서 학습시 정상 데이터를 사용하기 때문에 normal(N) Class만으로 학습 데이터를 구성하여 사용하고, N Class는 총 86,717개중 85,717개를 학습 데이터로 사용하고 1,000개는 검증 데이터로 사용함
- 정상과 부정맥으로 구성되는 검증데이터의 비율을 맞추기 위해 N클래스 1,000개와 S, V, F 클래스 각 330개, Q 클래스는 15개로 구성하며, 데이터의 비율이 맞지 않고 정상 데이터의 수가 많아진다면 성능평가의 결과가 편향될 가능성을 방지함

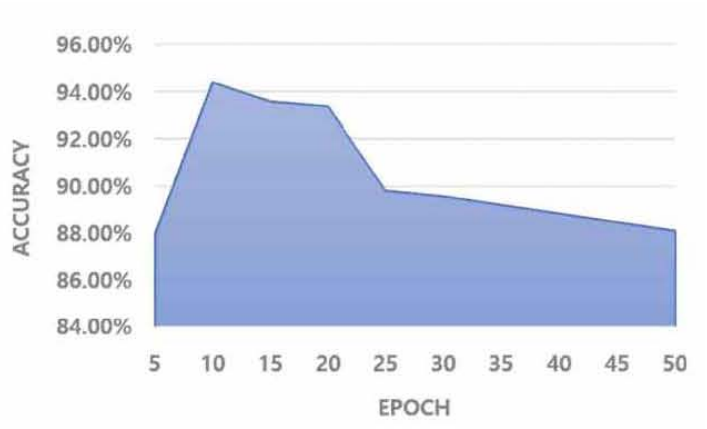
51) D. H. Shin, R. C. Park, K. Chung (2020) Decision Boundary-Based Anomaly Detection Model using Improved AnoGAN From ECG Data, IEEE Access, 8, 108664-108674.

■ **훈련 모델의 학습시 Epoch에 따른 모델의 정확도 평가**

- 정확도는 테스트에 사용된 전체 데이터의 수와 정확하게 분류된 데이터의 비율이며, 부류 불균형(class unbalanced) 상황에서 오류의 형태를 자세하게 반영하는 척도를 사용함. True Positive(TP), True Negative(TN), False Positive(FP), False Negative(FN)를 성능평가 요소로 활용함. Decision boundary를 200으로 설정하고 성능평가 실험을 진행함

$$Accuracy(\%) = \frac{TP + TN}{TP + TN + FP + FN}$$

• epoch에 따른 모델의 정확도 결과



- Epoch에 따른 제안한 모델은 10 Epoch에서 가장 높은 정확도를 가지며, 학습이 진행될수록 일반화 성능이 떨어진다고 판단되기 때문에 Epoch에 증가할수록 이상치 탐지에 대한 정확도가 낮게 평가됨

- 적합도 평가를 위해 머신러닝 기법인 KNN, One Class Support Vector Machine (OCSVM), Autoencoders (AE)와 딥러닝 모델인 CNN을 통한 ECG 분류 모델, 이상치 탐지에 사용되는 딥러닝 모델인 GaNomaly, 기존의 AnoGAN 모델을 비교함

- 적합도 검사에 사용한 지표는 Area Under a ROC Curve(AUC)와 F-measure를 사용하여 성능평가를 진행함. AUC는 Receiver Operating Characteristic curve(ROC) 커브의 밑면적의 넓이를 의미하며, AUC는 0~1의 값을 가지는데 1에 가까울 갈수록, 민감도는 높아지지만 특이도는 낮아지므로 성능이 좋음. Recall과 Precision간의 가장 안정적인 관계를 발견하여 정확성을 나타내는 척도인 F-measure를 사용함
- CNN과 GANomaly는 좋은 성능을 나타내지만, 학습 및 테스트 과정에서 라벨링된 데이터가 필요하다는 단점이 있고, 개선된 모델이 기존의 AnoGAN 보다 0.023, 0.0231 정도의 향상된 AUC 및 F-measure의 결과를 보이며, decision boundary를 객관적인 모델링 실험을 통해 좋은 성능을 보인다고 분석됨

■ **ECG Data에서 이상치 탐지 결과**

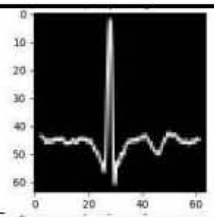
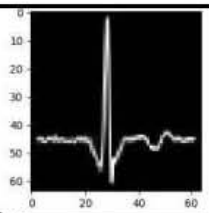
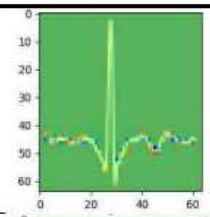
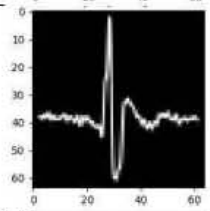
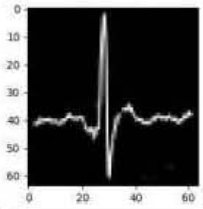
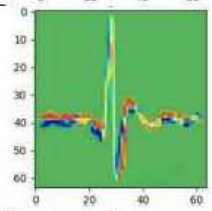
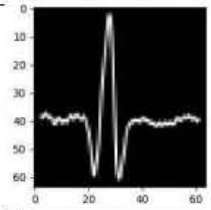
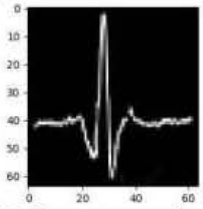
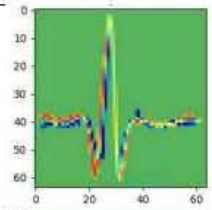
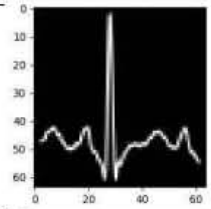
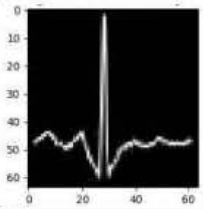
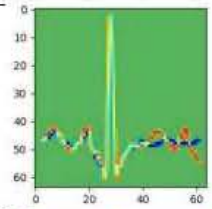
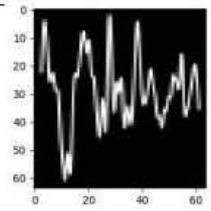
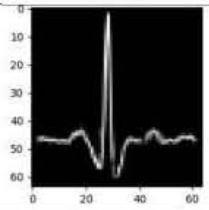
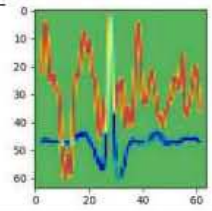
- 생성된 이미지와 쿼리 이미지의 차이를 비교하며 이상치 스코어를 통해서 decision boundary를 산정하고, 이상치 스코어는 두 가지 손실함수의 합으로 구함

$$Anomaly\ Score = (1-\lambda) \bullet RLoss + \lambda \bullet DLoss$$

- R Loss는 Generator의 Residual Loss이며, D Loss는 Discriminator Loss를 나타내고, λ는 기존 연구에서 진행한 성능평가에서 도출한 결과인 가장 효율적 성능을 보이는 0.1, nomaly score의 decision boundary는 200으로 설정하고 실험을 진행함

- 클래스별 Input Image(x)를 사용하며, Generator(G)에서 이미지(x)의 입력시 Latent Space(z)를 통해 생성되는 생성 이미지(Generated Image(G(z))와 x의 차이를 시각화하고, 이는 G(z)와 x의 픽셀별로 가중치 차이를 구함으로써 컬러 맵에 매핑하여 보여줌
- 생성 이미지와 x의 차이를 anomaly score(수치)도 나타내고, decision boundary를 통해 정상과 부정맥으로 분류함

• 생성 이미지와 AAMI Class별 Input 이미지의 비교를 통한 이상치 탐지 결과

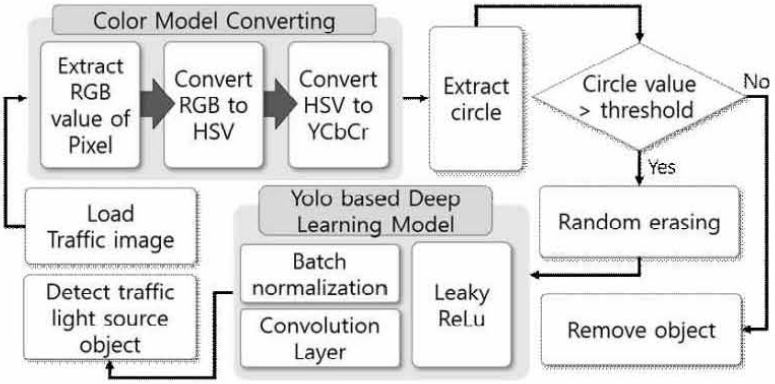
AAMI Class	Input Image(x)	Generated Image(G(z))	Anomaly Detection	Anomaly Score	Classification
N				170	Normal
S				454	Anomaly
V				497	Anomaly
F				351	Anomaly
Q				1401	Anomaly

- 실험 결과, N의 x를 입력할 경우는 Generator가 x의 확률분포와 유사한 G(z)를 생성하였기 때문에, 모델의 학습시 정상 데이터의 확률분포를 Latent Space(z)를 통해 근사함. x와 G(z)를 비교해 보면 시각적으로 큰 차이점이 없지만 S, V, F, Q의 x 입력 후 생성된 G(z)는 x의 확률분포와 다르게 생성되며, 정상 데이터의 확률분포만 z를 통해 학습하였고, 부정맥 데이터(S, V, F, Q)의 확률분포는 학습하지 않았기 때문에 다르게 생성한다고 분석됨

■ **색상모델 변화를 통한 딥러닝 기반의 광원객체 탐지의 성능평가**[52]

- 인식 정확도는 딥러닝 모델이 탐지한 광원 객체와 실제 값(Ground Truth)을 비교하여 얼마나 정확하게 인식했는지를 의미함
- 광원 객체가 뚜렷이 나타나지 않는 환경, 즉 야간 및 우천 등의 상황적 요인에 의해 발생하는 빛의 번짐과 백화현상은 RGB 모델의 인식 정확성을 하락시키는 주요 요인임. 이는 색상 단순화 과정을 거치는 HSV 모델을 통해 백화현상의 해결이 가능하므로 상대적으로 용이한 특징 추출이 가능함
- 야간도로 상황에서 색상 모델을 통한 광원 객체 탐지의 인식 정확도를 비교하기 위해 각기 다른 색상 모델에 따른 성능평가를 진행함

• 야간도로 상황에서 색상모델을 통한 광원 객체 탐지의 인식 정확도



	RGB	HSV	YCbCr	Ours
Accuracy of Detection	0.671	0.754	0.783	0.894

- 반면 RGB, HSV 생상 모델의 경우 빛의 번짐현상에 부정적인 성능 영향을 받음. 따라서 밝기와 색차 성분 분석을 사용하는 YCbCr 색상 모델을 통해 번짐 현상의 해결이 가능함. 그러나 기존의 색상 모델 모두 인식 정확도가 0.8 미만임
- 빛 번짐 및 백화현상의 해결이 가능한 모델을 조합함으로써 환경에 의해 발생하는 한계를 해결하고자 함. 이는 HSV와 YCbCr 모델의 합성의 경우 0.894로 기존 모델들의 인식 최고성능을 보장하며 향상된 결과를 나타냄. 이는 환경적 한계에 의해 광원 객체인 신호등을 용이하게 탐지하는 것이 가능함
- 교통영상을 데이터 수집시 발생하는 환경적 요인에 의한 영향을 최소화하고자 HSV 색상 모델과 YCbCr 색상모델을 조합함. 이는 밝기 성분 분석을 통해 빛 번짐 현상과 백화현상에 의한 탐지 인식률 저하를 방지함. 딥러닝 모델을 기반으로 무작위 삭제 방법을 적용한 Yolo 모델을 학습함
- 작은 객체의 탐지가 가능한 CSPDarknet53기반 모델을 이용하여 클래스 불균형 문제를 해결하였고, 효율적인 교통 영상에서 광원 객체 탐지를 통해서 자율주행의 핵심 부분인 신호등 인식 단계의 인식률을 높이고 주간 도로 상황뿐만 아니라 야간 도로상황에서도 적용이 가능함

■ **교통사고 토픽 관련 트랜잭션에서 연관규칙의 생성 속도 비교**[53]

- 뉴스 스트림 데이터에서 크롤링을 통해 1700개의 교통사고 관련 데이터를 수집하여 코퍼스로 구성함
- 수집된 코퍼스는 분석의 질을 향상시키기 위해 데이터 정제과정을 진행하여 결측값 처리와 이상값 처리, 불필요한 데이터를 제거하여 데이터의 차원 축소

52) 강지수, 심세은, 조선문, 정경용, 교통 영상 빅데이터 처리를 위한 Yolo 기반 광원 객체 탐지, 융합정보논문지, 제10권, 제8호, pp. 40-46, 2020.

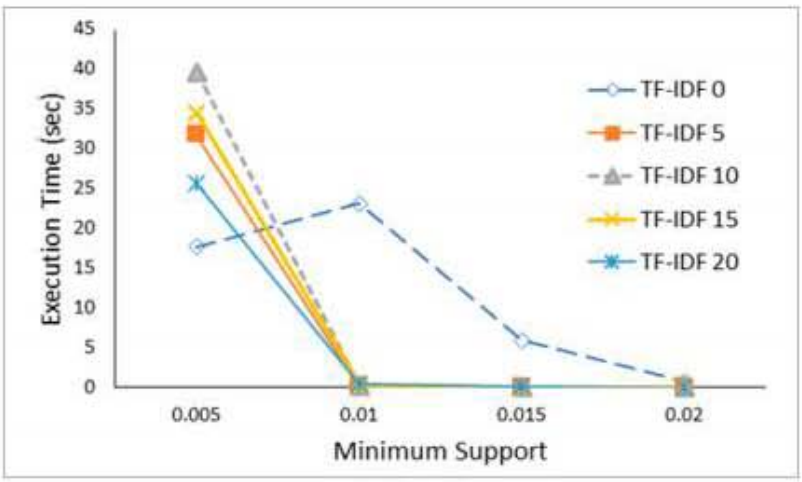
53) H. J. Kim, J. W. Baek, K. Chung (2020) Optimization of Associative Knowledge Graph using TF-IDF based Ranking Score, Applied Science, 10(13), 4590-4910.

- 형태소분석을 통해 교통사고 토픽을 추출하고 TF-IDF 가중치를 이용하여 트랜잭션으로 구성
- 연관규칙 알고리즘의 성능을 비교하기 위하여 FP-Tree와 Apriori 알고리즘의 결과 데이터를 각각 생성

▪ 최소 support와 confidence의 독립적인 변화에 따른 연관규칙의 생성속도 비교

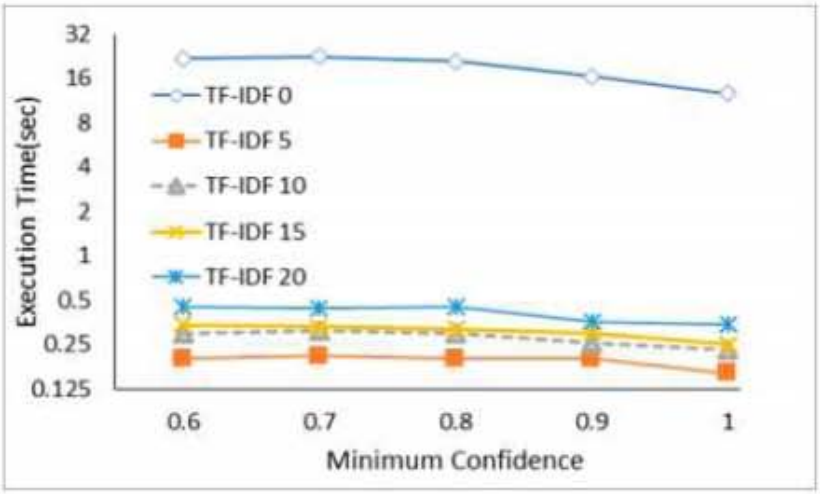
- TF-IDF 가중치를 기반으로 단어 순위를 이용한 연관단어 그래프 생성과 TF-IDF를 적용하지 않은 연관단어 그래프 생성의 성능을 평가하기 위해서 연관규칙의 생성속도 비교
- TF-IDF의 상위 5%, 10%, 15%, 20%에 따른 생성속도와 개수를 비교하여 실시간으로 생성되는 스트림 데이터의 특성을 고려하여 지식 생성에 적합한 지지도와 신뢰도 판단

• TF-IDF와 최소지지도 변화에 따른 연관규칙의 생성속도 비교 결과



- 0.1로 고정된 confidence 값이 있는 상태에서 support 값이 0.005일 때, 상위 15%, 20%의 단어를 적용한 상태에서 TF-IDF를 적용하지 않은 모델 대비 연관규칙을 생성하는데 약 16.7, 8초 더 소요
- support를 0.01 이상으로 설정하였을 때, 22초 이상의 차이로 연관규칙의 생성속도가 대폭 감소하고, 0.01이상의 모든 값에서 상위 TF-IDF 단어를 적용한 모델이 연관규칙 생성시간이 감소함
- support는 0.01로 고정을 함에 따라 confidence에 변화를 줄 경우의 연관규칙의 생성속도를 비교함

• TF-IDF와 신뢰도 변화에 따른 연관규칙의 생성속도 비교 결과



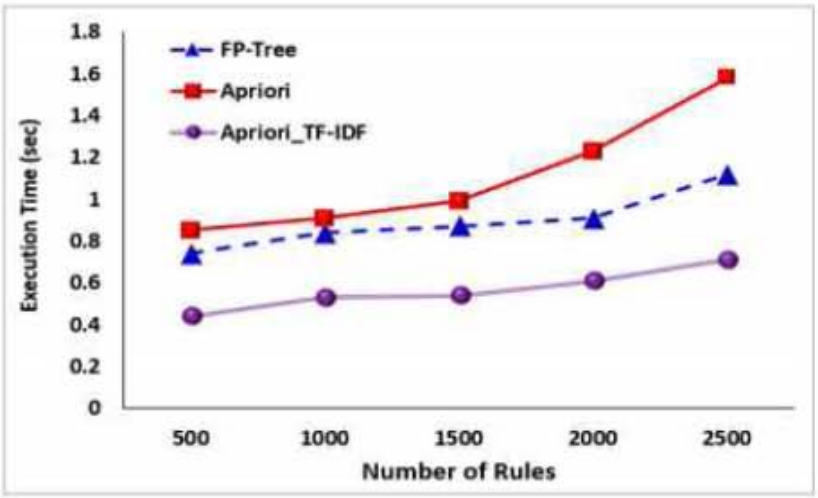
- confidence에서 TF-IDF의 유무에 따라 연관규칙의 생성속도가 12초, 약 44배 이상의 차이 보임
- TF-IDF를 적용한 모델의 경우 confidence에서 1초를 넘지 않으며 기존의 transaction에 있어야할 불필요한 단어의 수가 감소했기 때문에 신뢰성 있는 연관규칙의 생성이 기존의 연관규칙 알고리즘보다 규칙의 생성속도가 빠르다는 것을 보여줌

■ **지지도와 신뢰도, 향상도를 이용하여 Apriori 알고리즘과 FP-tree 알고리즘을 비교**

- Apriori 알고리즘은 모든 트랜잭션에서 단어들에 대한 연관규칙을 생성하고, 지지도와 신뢰도를 이용하여 가지치기를 진행하지만 규칙생성 속도의 성능이 낮음
- FP-Tree는 빈발항목 패턴에서 linked list를 사용하여 패턴을 생성함으로써 빈발 패턴을 마이닝하여 효율적이고 확장이 가능하며 apriori 알고리즘보다 탐색 속도가 빠른 장점이 있음

- 생성된 rule의 개수를 제한시켜 출력시킨 후, 알고리즘별로 Support, Confidence, Lift에 대하여 평균을 계산하여 FP-tree 연관규칙 알고리즘, Apriori 알고리즘과 개선된 Apriori 연관규칙 알고리즘에서 추출한 연관규칙 생성속도와 유용성 비교하기 위해 rule의 개수는 500, 1000, 1500, 2000, 2500개로 각각 제한한 후, 성능평가 진행함

• FP-Tree, Apriori, Apriori_TF-IDF 알고리즘의 규칙 생성 속도 비교



- Apriori-TF-IDF 알고리즘이 기존의 알고리즘보다 0.4~0.8초 빠르게 규칙을 생성하며, 객관적 평가를 위해 지지도, 신뢰도, 향상도를 사용함

• 지지도와 신뢰도 평가 결과

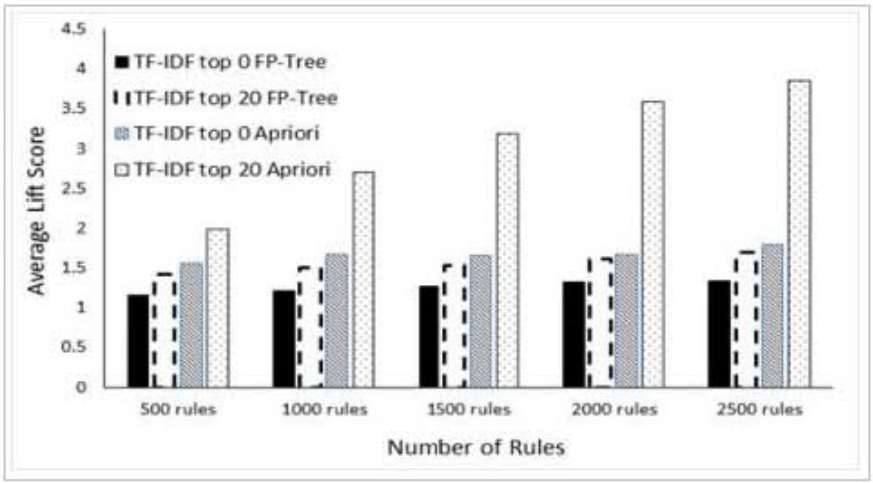
Support	500 Rules	1000 Rules	1500 Rules	2000 Rules	2500 Rules
TF-IDF top 0 FP-Tree	0.09979	0.08379	0.07304	0.06389	0.06087
TF-IDF top 20 FP-Tree	0.05093	0.04188	0.03635	0.03278	0.03036
TF-IDF top 0 Apriori	0.1404	0.11343	0.09818	0.09097	0.08495
TF-IDF top 20 Apriori	0.06794	0.05186	0.04522	0.04133	0.03903
Confidence	**500 Rules**	**1000 Rules**	**1500 Rules**	**2000 Rules**	**2500 Rules**
TF-IDF top 0 FP-Tree	0.17034	0.17132	0.17251	0.17189	0.17189
TF-IDF top 20 FP-Tree	0.16144	0.16273	0.16123	0.16078	0.16114
TF-IDF top 0 Apriori	0.4745	0.5077	0.4941	0.4877	0.4902
TF-IDF top 20 Apriori	0.3793	0.4138	0.4398	0.4559	0.4653

- TF-IDF Ranking Score를 적용하지 않은 FP-Tree와 Apriori가 average support값에 대하여 ranking을 적용한 알고리즘이 더 높은 점수를 나타냄
- confidence 값에서는 rank를 적용하지 않은 Apriori, 상위 20% TF-IDF Rank를 적용한 Apriori, rank를 적용하지 않은 FP-Tree, 상위 20% TF-IDF Rank를 적용한 FP-Tree 순으로 높은 점수를 나타냄
- 향상도를 이용하여 규칙의 효용성 검증

• 규칙 수의 변화에 따른 알고리즘의 향상도 평가 결과

Lift	500 Rules	1000 Rules	1500 Rules	2000 Rules	2500 Rules
TF-IDF top 0 FP-Tree	1.15356	1.2108	1.2697	1.31759	1.34029
TF-IDF top 20 FP-Tree	1.42306	1.50636	1.52958	1.62002	1.69837
TF-IDF top 0 Apriori	1.5555	1.6721	1.6558	1.6719	1.794
TF-IDF top 20 Apriori	1.991	2.7055	3.1918	3.5879	3.8551

• 규칙 수에 따른 연관 알고리즘의 평균 향상도 점수



- Lift을 이용한 평가 결과 기존의 연관규칙 알고리즘 대비 최대 두 배 이상의 성능을 보임
- rules의 값을 늘릴수록 TF-IDF Ranking을 적용한 알고리즘이 더 높은 성능을 보였기 때문에 실시간으로 수집되는 스트림 뉴스 코퍼스에서는 연관규칙의 수가 증가하여도 규칙의 생성속도와 유용성 및 효용성을 향상시키기 때문에 지식 그래프의 최적화로 유의미한 정보를 실시간으로 추출이 가능함

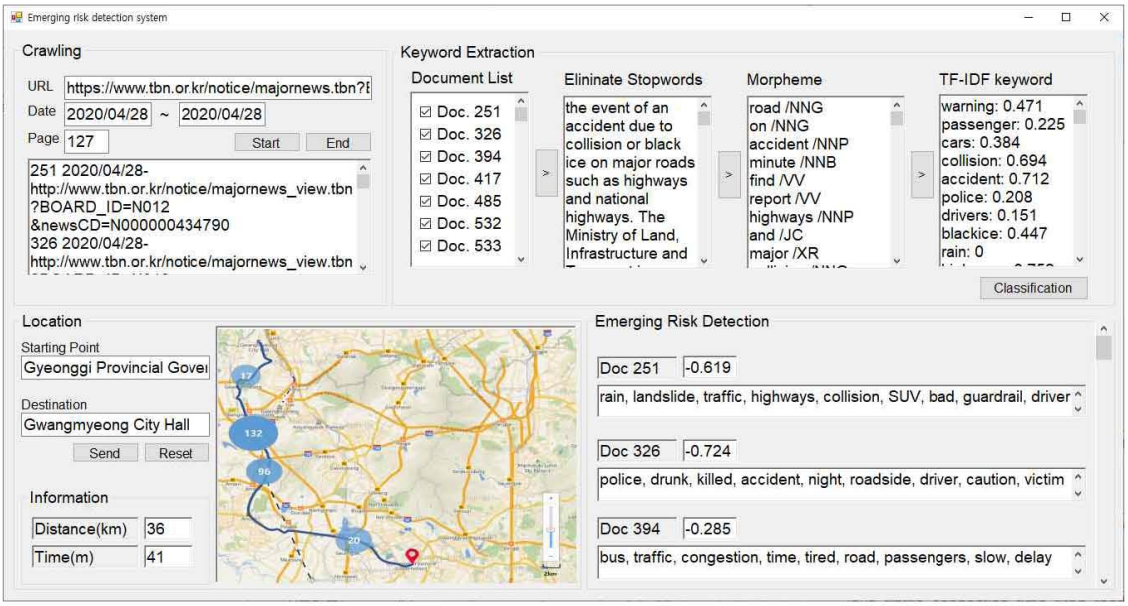
■ 이머징 리스크 탐지를 위한 워드 임베딩 기반 교통문서 분류에 따른 성능평가[64]

- 감성 유사도 가중치 기반 분류 모델을 교통상황에 적용하여 사용자 경로를 기반으로 시스템 구축하여, 사용자의 경로에 존재하는 이머징 리스크를 탐지하고 리스크에 대한 정보를 제공 함으로써, 교통위험 경로의 전체적인 이머징 리스크 정보를 사용자별로 제공함
- 이머징 리스크 탐지 시스템은 크롤링, 키워드 추출, 위치 및 이머징 리스크로 구성됨. 먼저, url, 날짜, 페이지를 설정한 뒤 크롤링을 시작하고, 문서가 선택되면 불용어가 제거하여 진행함
- 불용어가 처리된 문서에 대해 형태소분석 과정을 진행하고, 전처리된 문서에서 TF-IDF 값을 얻은 후 SSW 기반의 SVM을 이용하여 분류를 수행함. 위치 부분에서 시스템은 사용자의 출발 및 목적지 지점을 수신하고 경로와 관련된 리스크 문서의 수를 지도에 표시하고, 문서에서 추출한 이머징 리스크 키워드를 기반으로 각 경로에 대한 간단한 정보를 사용자에게 제공할 수 있음

- 시스템을 통해 Gyeon Provincial Government에서 Gwangmyeong City Hall까지의 경로에 따른 emerging risk 탐지를 나타내며, 지도에서는 4개의 구간에서 총 265개의 emerging risk 문서가 탐지됨. 그 중 수원광명고속도로 구간은 132개의 문서가 탐지되며 가장 위험성이 높다고 판단되며, 추출된 문서들에서 감지된 키워드가 표시되므로 모든 구간의 위험 요소들을 확인할 수 있고, 경로의 전체적인 리스크와 이머징 리스크 정보를 사용자에게 제공함으로써 교통위험의 대비가 가능함
- 감성 유사도 가중치를 이용한 교통문서 분류 모델을 구현하고 성능평가를 수행하기 위해 TBN의 뉴스 데이터를 크롤링하며, 문서는 총 1,400개를 수집하여 70%의 훈련데이터와 30%의 테스트 데이터로 분할하여 사용함

64) M. J. Kim, J. S. Kang, K. Chung (2020) Word-Embedding-Based Traffic Document Classification Model for Detecting Emerging Risks Using Sentiment Similarity Weight, IEEE Access, 8, 183983-183994.

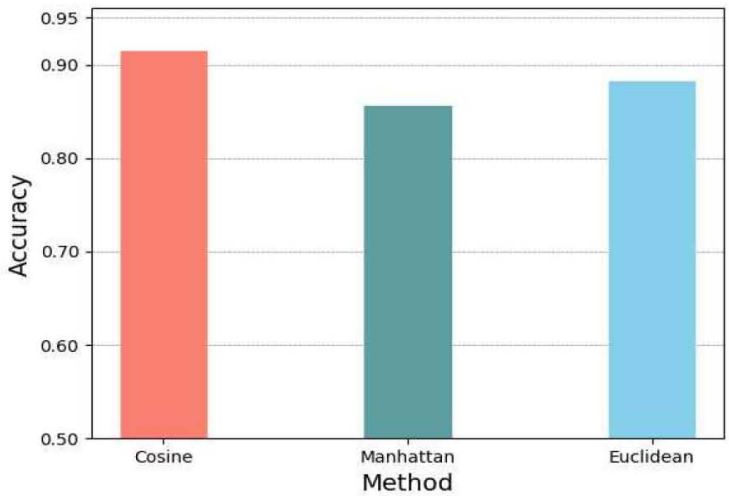
• 감성 유사도 가중치 기반 분류 모델을 적용한 사용자 경로 기반 이머징 리스크 탐지 시스템



■ **키워드 간의 유사도를 측정할 때 가장 우수한 유사도 평가 방식을 판단**

- 단어 간의 유사도 측정 방식에 따른 모델의 정확도를 비교하여 평가하기 위해 벡터 공간에서 코사인 유사도를 이용하여 단어 사이의 유사도 값을 측정함
- 사용한 코사인 유사도 방식이 가장 우수한 유사도 측정 방식임을 증명하기 위해 코사인 유사도와 맨하탄 유사도, 유클리디언 유사도를 적용하여 이에 따른 문서 분류의 정확도를 비교함
- 세로축은 accuracy를 의미하며 가로축은 코사인, 맨하탄, 유클리디언의 유사도 방식을 나타냄

• 유사도 방식에 따른 정확도 비교 결과

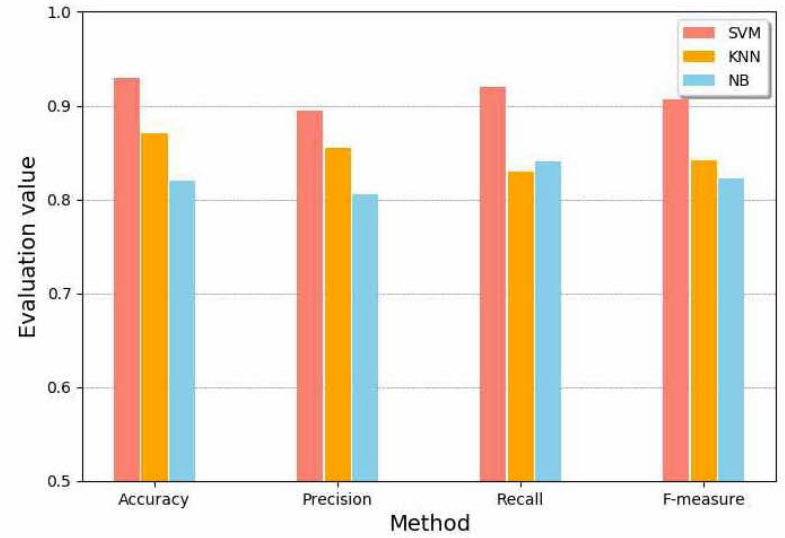


- 정확도 평가 결과, 코사인 유사도가 다른 방식보다 각각 0.059, 0.033만큼 우수하게 평가되며, 코사인 유사도는 벡터 사이의 방향성을 통해 다차원의 공간에서 유사도를 측정하기 때문에 문서 안에서의 전체적인 키워드의 맥락을 고려하여 측정 가능함
- 맨하탄과 유클리디안 등의 거리 기반 유사도 방식은 단순히 단어의 빈도만을 고려하여 같은 맥락의 문서들을 다르다고 판단하기 때문에 자연어 처리에 적합하지 않은 한계가 존재함. 각도 기반의 코사인 유사도와 비교하여 거리 기반의 맨하탄과 유클리디언 유사도 방법이 낮은 정확도를 보임
- 단어들의 문맥적 의미를 통해 워드 임베딩 과정이 이루어지기 때문에 각 단어의 방향성이 고려되는 코사인 유사도 방법이 적합함. 이는 코사인 유사도를 사용하여 각 키워드간의 유사성에 대해 의미적인 추론을 진행하였음

■ **감성 유사도 가중치를 적용하여 이머징 리스크 탐지에 대한 적절한 분류 모델의 평가**

- 사용한 SVM 분류 모델을 비롯한 KNN, Naive Bayes 분류 모델과 emerging risk 문서 분류의 성능을 비교하여, 가장 우수한 분류 모델을 선정하기 위해 다양한 분류 모델에 SSW를 적용하여 분류의 성능을 평가함
- 분류 모델은 사용한 SVM 외에도 KNN 모델과 Naive Bayes를 선정하여 accuracy, precision, recall, F-measure를 측정하여 비교함. precision은 모델이 이머징 리스크로 예측한 것 중 실제 리스크에 해당되는 확률을 의미하고, recall은 모든 이머징 리스크 문서 중에서 모델이 정확하게 분류하는 확률을 의미함
- 데이터가 불균형한 상황에서는 정확도의 척도가 불안정해지기 때문에 정밀도와 재현률을 기반으로 하는 F-measure를 accuracy와 함께 성능평가 척도로 채택하며, F-measure는 precision과 recall의 조화평균이며, 분류의 정확도를 결정하는데 사용함

• 각 분류 모델에 대한 accuracy, precision, recall, F-measure의 비교 결과



- 적용했을 때 accuracy, precision, recall, F-measure 모두 SVM 분류 모델이 가장 좋은 성능을 보였음. SVM 분류 모델이 적은 데이터 집합에도 좋은 성능을 보이며 고차원의 데이터에 적합하기 때문에 상대적으로 높은 성능을 나타내는 것으로 판단됨

- 학습데이터는 여러 개의 키워드를 포함하는 고차원 속성으로 이루어져있으며, 약 1400개의 소규모 데이터이기 때문에 SVM 분류 모델이 적합하지만, Naive Bayes 분류 모델은 키워드간의 중요도를 고려하지 않고 각각의 독립적인 요소로 판단하기 때문에 성능지표가 낮게 평가됨
- KNN은 다소 높은 정확도를 보이지만, 키워드의 유사도를 계산해야하기 때문에 연산시간과 비용이 과도하게 소모되어 적합하지 않기 때문에 SVM 모델이 고차원의 키워드를 포함한 텍스트 마이닝에 가장 적합한 분류기임

■ **감성 유사도 가중치를 적용한 교통위험 문서의 탐지에 대한 유용성 평가**

- SVM 이진 분류기에 감성 유사도 가중치를 적용한 모델과 적용하지 않은 기존 모델의 비교를 통해 평가를 진행하였고, 가장 적합하다고 판단된 SVM 분류기에 개발한 감성 유사도 가중치의 적용 여부에 따른 accuracy, precision, recall, F-measure를 비교를 진행함
- 감성 유사도 가중치가 이머징 리스크 탐지에 유의미한 가중치로써 적용된다는 것을 증명하고, 성능평가 결과에 따른 ROC curve 그래프를 통해 시각화를 진행함

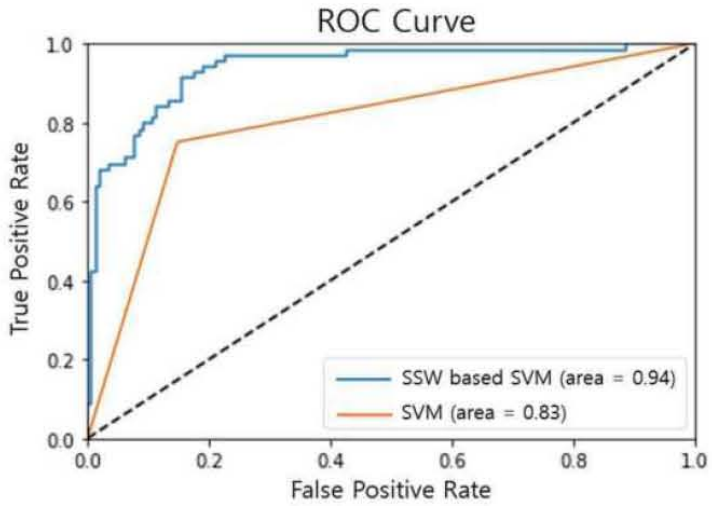
- 감성 유사도 가중치 기반 분류 모델의 accuracy가 기존의 분류모델에 비해 0.118 만큼 우수하게 평가되고, F-measure 평가에서는 0.085만큼 높게 평가됨. 감성 유사도 가중치를 통해 감성사전에 없는 단어를 효과적으로 대치하여, 문서 내에서의 단어의 중요성을 고려하여 감성사전에 단어가 존재하지 않더라도 단어 간의 유사도를 고려한 대치 방법으로 인해 좋은 성능을 보임

• SSW를 적용한 SVM 모델과 기존 모델의 성능 비교 결과

	Evaluation Method	Value
SSW-based SVM	Accuracy	0.930
	Precision	0.895
	Recall	0.920
	F-measure	0.907
SVM	Accuracy	0.812
	Precision	0.805
	Recall	0.840
	F-measure	0.822

- 감성 유사도 가중치를 적용하지 않은 일반 SVM 분류 모델은 감성사전에 존재하지 않는 단어들을 모두 배제하고 분류를 수행하기 때문에 단어들의 긍정/부정을 고려하지 않아 낮게 나타남
- 감성 유사도 가중치를 이용하여 감성 사전에 존재하지 않는 단어와 존재하는 단어 모두 효과적으로 유의미하게 가중치를 부여하는 것이 가능하다고 판단
- 성능평가 결과를 시각화하기 위해 ROC(Receiver Operating Characteristic) curve를 사용하고, ROC curve는 x축을 FPR(False Positive Rate), y축은 TPR(True Positive Rate)로 한 그래프를 의미함. AUC(Area Under Curve)는 ROC curve의 아래쪽 면적의 크기를 의미하며 0에서 1 사이의 값을 지니며, AUC의 값은 1에 근접할수록 좋은 성능의 모델이라고 판단할 수 있음
- TP는 분류기가 이미징 리스크 문서를 리스크라고 판단했음을 나타내고, FN은 이미징 리스크 문서를 비-리스크라고 판단했음을 나타냄. FP는 non-emerging risk 문서를 emerging risk로 판단했음을 나타내고, TN은 non-emerging risk를 non-emerging risk로 판단했음을 나타냄

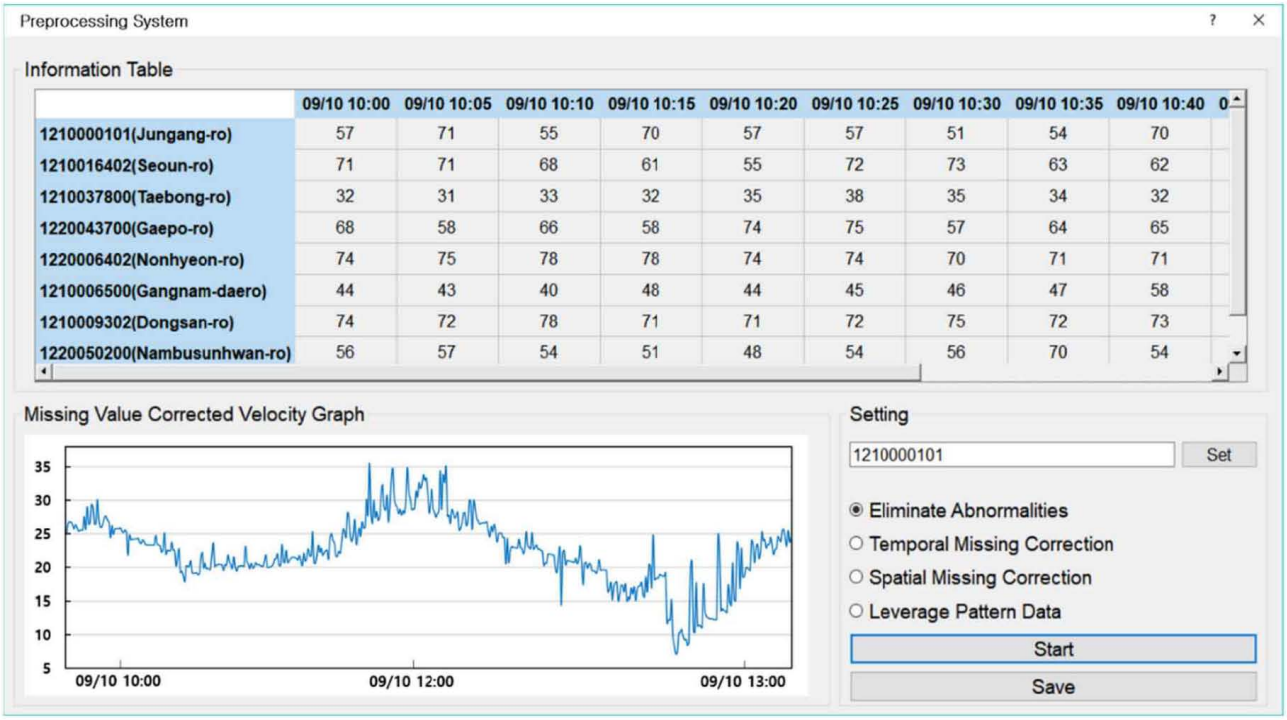
• SSW 적용 여부에 따른 ROC curve 비교



- (a)는 SSW를 적용한 SVM 분류 모델을 나타내며, (b)는 SSW를 적용하지 않은 기존 SVM 분류 모델의 ROC curve를 나타냄. ROC curve 시각화 결과 각 모델의 AUC(Area Under Curve)는 각각 0.94, 0.83으로 SSW 적용 모델의 분류성능이 더 우수함. SSW를 이용한 모델에서 유의미한 분류를 통해 교통 이머징 리스크를 감지하는 것이 가능함

■ 교통 혼잡도를 이용한 사고 위험도 예측모형에 대한 성능평가[55]

• LSTM 모델 기반 교통 혼잡도 예측을 위한 전처리 시스템



• LSTM 모델 기반 교통 혼잡도 예측 시스템



- 교통 데이터를 전처리하는 과정과 교통 혼잡도 예측 모델을 사용자가 사용할 수 있도록 시스템을 개발함
- 예측 시스템의 선택 창에서 구간을 선택하고, LINK overview 창에서는 선택한 구간에 대한 설명(LINK_ID, LINK_NAME, Velocity), Data collection 창에서는 선택한 구간의 날짜별 속도 수집현황을 보여줌. [Status]에서 5분후의 예측을 선택하고 [Predict] 버튼을 누르면 서울특별시 강남구의 전반적인 혼잡도 결과를 보여주며 혼잡도는 원활(초록), 복잡(주황), 매우 복잡(빨강)으로 표시된 결과를 보임
- 강남구 전체 구간별 속도데이터를 나타낸 것이며, 하단에 LINK_ID를 입력하고 필요한 전처리 과정(이상치 제거, 공간적 결측 보정, 시간적 결측 보정, 패턴데이터 활용)을 선택하고 [Set] 버튼을 눌러 전처리를 진행

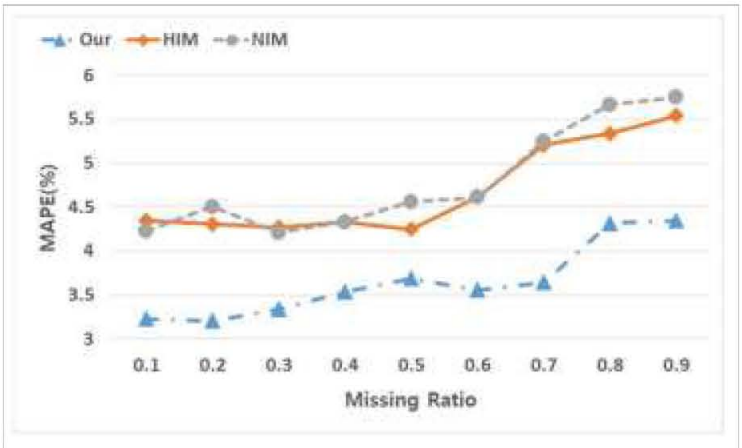
55) D. H. Shin, K. Chung, R. C. Park (2020) Prediction of Traffic Congestion Based on LSTM Through Correction of Missing Temporal and Spatial Data, IEEE Access, 8, 150784-150796.

함. 전처리가 완료된 구간 데이터는 그래프로 나타내고, 30Km/h 이상(원할), 15Km/h~30Km/h(복잡), 15Km/h 미만(매우 복잡) 기준으로 구분하여 결과를 판단함

■ 데이터 결측에 따른 예측 모델 성능평가

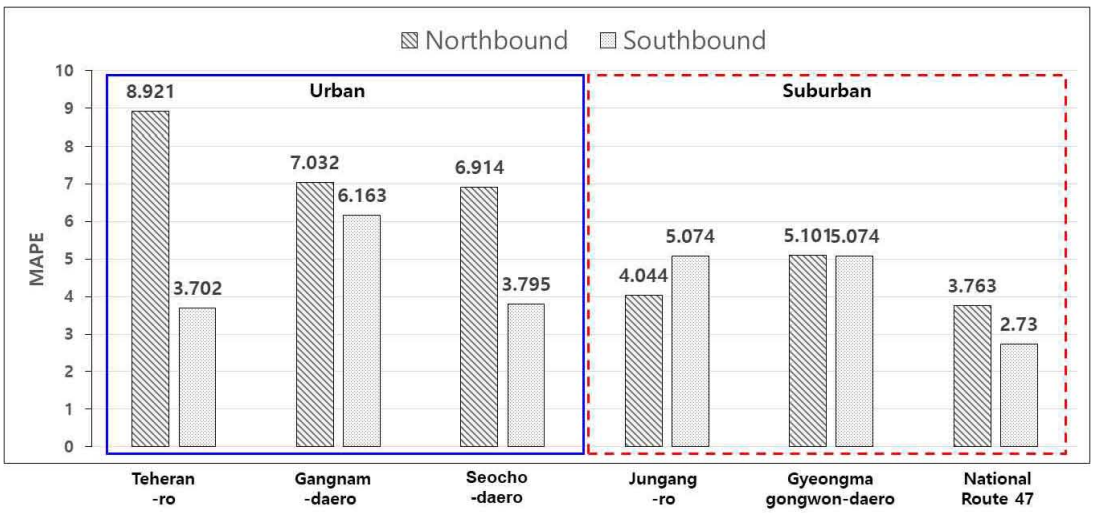
- 데이터의 이상치 및 결측으로 인한 모델의 성능 하락을 방지하기 위해 중위수 절대 편차를 이용한 이상치 제거 및 시공간적 결측 보정을 통해 전처리를 진행. 결측률에 따른 모델 성능을 분석하기 위해 인위적으로 데이터 결측률을 0 ~ 40% 범위의 데이터로 생성. 생성된 데이터를 통해 예측을 진행
- 데이터 결측률이 0%인 데이터가 평균 5%의 MAPE 지수로 우수한 성능을 나타냄. 결측률이 40%인 데이터는 수집한 데이터 그대로를 사용하여 진행한 결과인데 이는 성능의 편차가 크고 평균 38%의 MAPE 지수로 좋지 못한 성능을 나타냄

● 데이터 결측률에 따른 예측 모델의 성능평가 결과



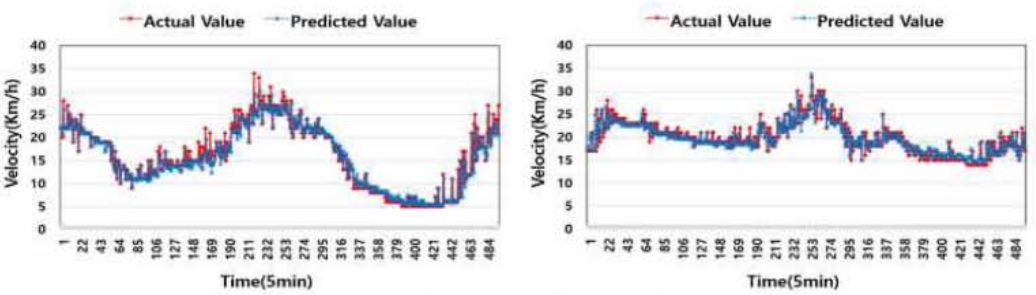
- 모델의 성능평가와 손실함수로 절대 평균 백분율 편차(MAPE: Mean Absolute Percentage Error)을 사용한다. MAPE는 크기 의존적 에러의 단점을 보안하기 위한 모델이며, 실제값과 예측값의 오차 절댓값의 평균을 의하며, 정체구간과 같은 저속에서 작은 값의 예측에 민감한 MAPE의 특성이 있어 손실함수로 사용
- 연속류 구간은 외부 영향이 없어 교통 흐름이 통제되거나 변형이 없는 구간이고, 단속류 구간은 신호등과 같은 통제구조물 혹은 시설에 의해 교통 흐름이 단절되는 구간으로 교차로 및 간선도로 시설이 있는 구간을 뜻함. 연속류 구간은 주로 고속도로가 있는 도심외곽지이며 도심지는 신호등 및 교통통제시설로 인하여 주로 단속류 구간이 주를 이룸

● 도심 외곽 지역 및 도심지 MAPE 결과

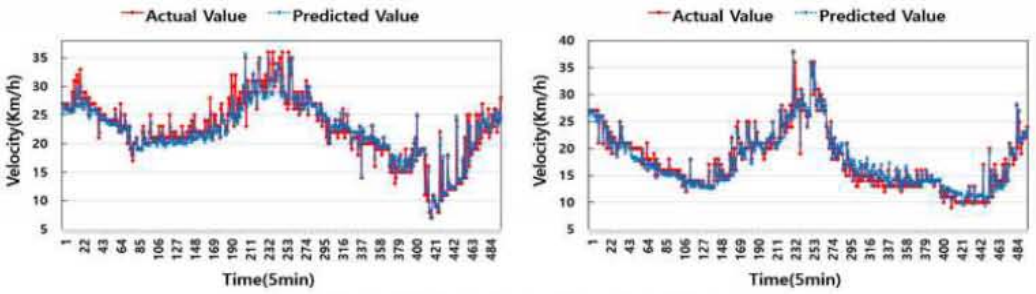


- 도심 외곽지역을 대상으로 세 개의 구간을 추출하여 상, 하행선 속도를 예측한 결과 그래프를 보여주며 평균 MAPE가 약 4.297%의 결과가 도출되었고, 도심지를 대상으로 도심 외곽지역과 마찬가지로 세 개의 구간의 상, 하행선 속도를 예측한 결과이며 평균 MAPE가 약 6.087%로 결과를 도출함

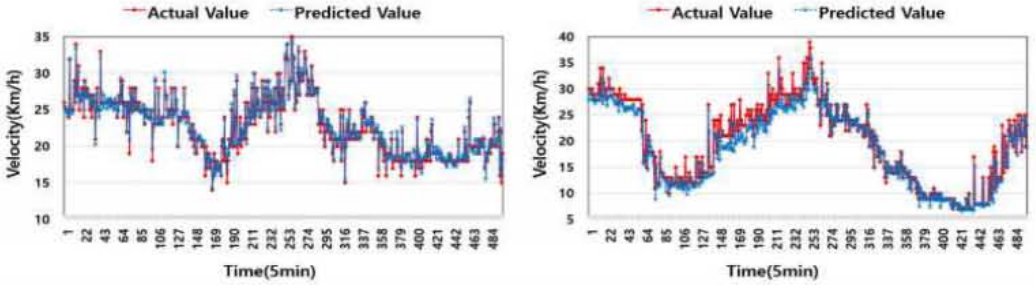
• 도심지역 분석 결과



(a) Upper region of Seocho-daero (b) Lower region of Seocho-daero

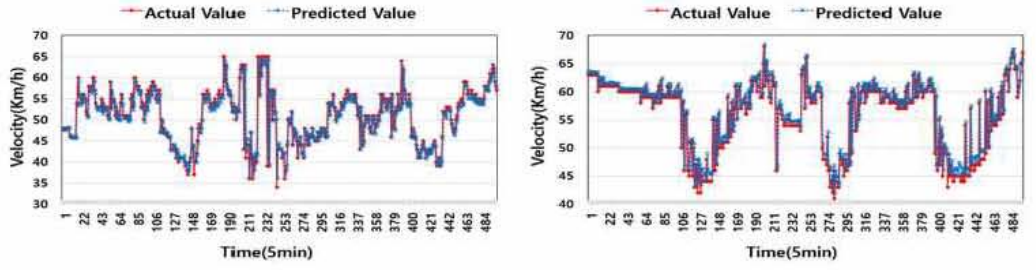


(a) Upper region of Gangnam-daero (b) Lower region of Gangnam-daero

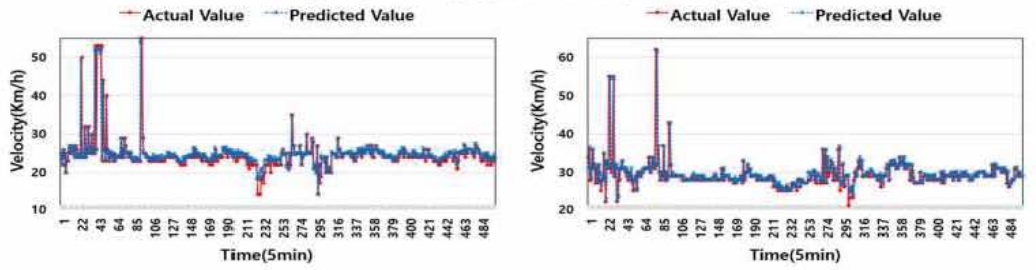


(a)Upper region of Teheran-ro(b) Lower region of Teheran-ro

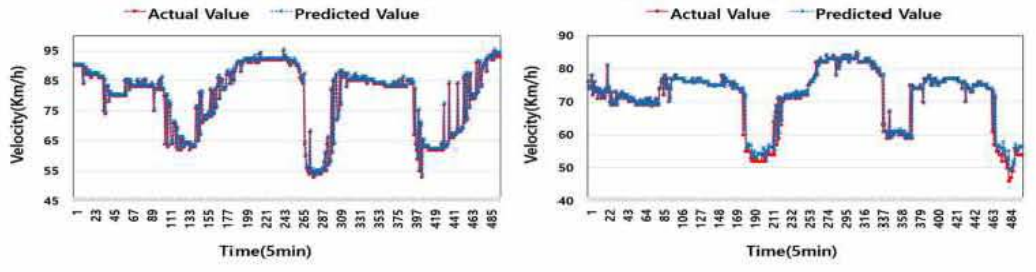
• 도심 외곽 지역 분석 결과



(a)Upper region of Jungang-ro (b) Lower region of Jungang-ro



(a)Upper region of Gyeongmagongwon-daero (b)Lower region of Gyeongmagongwon-daero



(a)Upper region of National route No. 47 (b)Lower region of National route No. 47

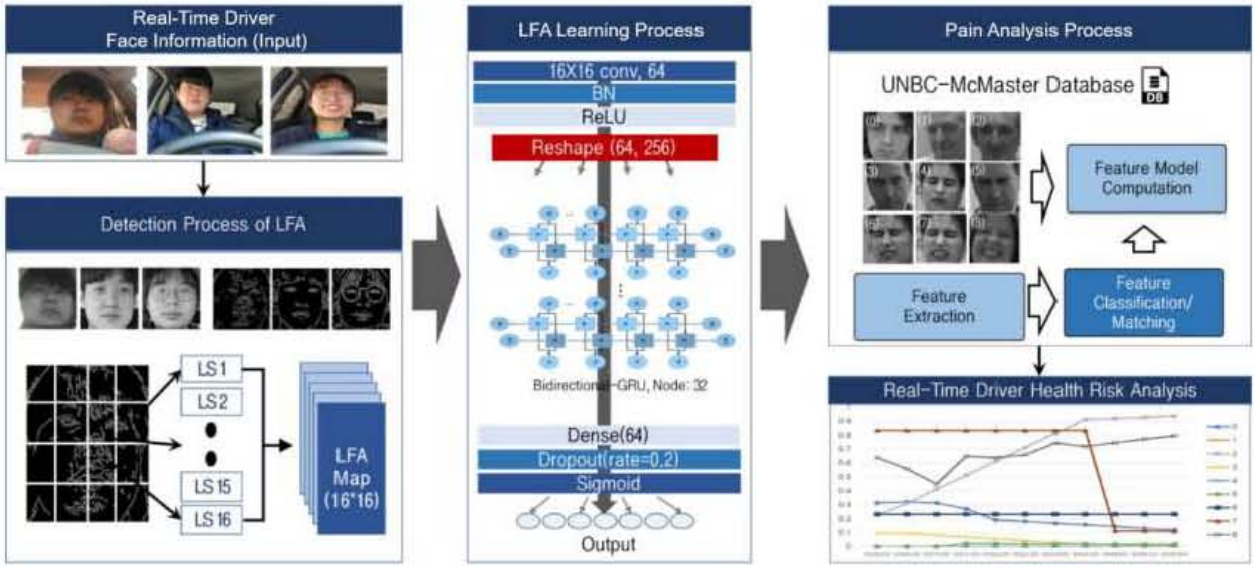
- 결과를 비교하여 보았을 때 도심지가 도심 외곽지역보다 다소 정확도가 떨어지는 것으로 결과가 도출. 이와 같은 결과가 도출된 이유를 분석한 결과, 도심 외곽과 같은 구간은 주변에 건물이 적고 신호등이 없으며 도심지 도로보다 제한속도가 높다는 특징이 있음
- 그에 반해 도심지 도로의 특징은 주변에 건물이 다수 존재하며 운전자 외 유동인구의 영향이 크고 교차로마다 신호등이 존재하여 교통 흐름에 차질이 생기는 변수가 많기 때문에 예측에 어려움이 크다고 판단됨

- LSTM 모델에 대한 신뢰성을 확인을 위해 기존의 연구에 사용된 예측 모델 SVR, KNN, CNN, MLP LSTM 모델을 비교에 사용. 기존 연구에 사용된 SVR, KNN, CNN의 MAPE 값은 각각 12.8%, 22%, 8.3% 이다. CNN을 사용한 연구는 비교적 좋은 결과를 가짐. 연구에서 사용한 LSTM 모델이 MAPE 값 5.187%로 기존의 연구들보다 우수한 것을 알 수 있음

■ 운전자 안면인식 기반 사고위험도 예측 모형 평가[56]

- LFA-CRNN 모델 기반 얼굴인식(고통 및 비-고통 표정)의 성능평가를 위해 UNBC-McMaster 데이터베이스를 사용하였으며, AlexNet과 CRNN 모델과 성능 비교를 수행함. UNBC-McMaster 데이터베이스는 Prkachin and Solomon Pain Intensity Scale (PSPI) score를 통해 고통을 9(0~8) 단계로 분류하였으며, 129명(63명의 남자와 66명의 여자)으로 구성됨

  • 운전자의 상태 분석 및 사고위험도 예측의 평가 과정[57][58]



- UNBC-McMaster 데이터를 기반으로 학습률 및 loss 측정 실험으로, 전처리(얼굴검출 및 윤곽선 추출) 및 LFA 변환과정을 통해 산출된 데이터를 LFA-CRNN의 입력으로 사용했으며, 성능 비교를 위한 CRNN, AlexNet은 얼굴검출 과정을 통해 산출된 데이터를 사용함

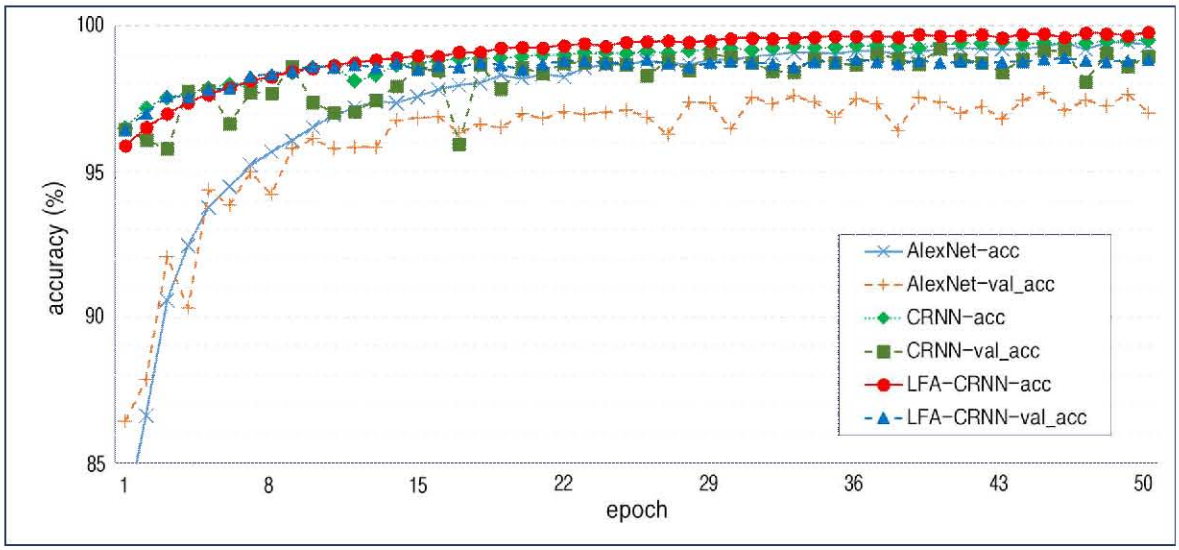
- UNBC-McMaster 데이터베이스에 대한 20%의 데이터를 실험의 테스트 데이터로 분류, 나머지 80% 중 10%를 검증데이터로 활용하여 실험을 진행하고, 데이터의 분류 시 특정 클래스에 편중되지 않도록 하기 위해 각 클래스별로 퍼센트를 지정하여 분류함

56) C. M. Kim, E. J. Hong, K. Chung, R. C. Park (2020) Driver Facial Expression Analysis Using LFA-CRNN-Based Feature Extraction for Health-Risk Decisions, Applied Sciences, 10(8), 2956-2976.
57) 특허출원, 특징 추출 기반의 LFA-CRNN을 이용한 운전자 표정 분석 장치 및 그 방법, 출원번호 10-2021-0109154, 2021.08.19.
58) 신동훈, 백지원, 박찬홍, 정경용, 도로 CCTV 데이터를 활용한 딥러닝 기반 차량 이상 감지, 한국융합학회논문지, 12(2), 1-6, 2021.

■ UNBC-McMaster 데이터베이스를 이용한 학습율 및 손실 측정 결과

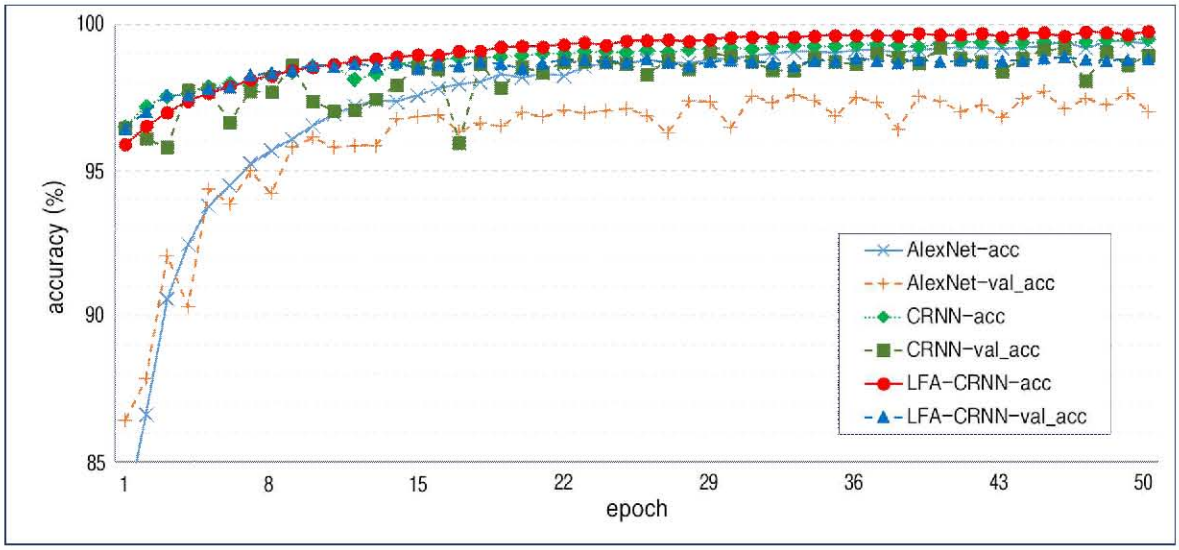
- LFA-CRNN의 학습률이 가장 높으며, AlexNet, CRNN의 순으로 높은 순을 보임
- AlexNet은 학습률과 검증 데이터 간에 격차가 많이 발생되는 것을 볼 수 있으며, CRNN의 학습률은 지속적인 상승을 보이는데 반해 검증 데이터를 이용한 검증률은 과적합으로 인하여 한시적으로 검증률이 떨어지는 것을 확인할 수 있음
- LFA-CRNN은 다소 학습률과 검증률에 대한 격차가 발생되었지만 심하지 않으며, 검증률이 한시적으로 떨어지는 문제가 발생되지 않는 것으로 보아 학습에 대한 과적합이 발생되지 않는 것을 확인할 수 있음

• UNBC-McMaster 데이터베이스를 이용한 학습율



- AlexNet은 학습 데이터에 대한 loss와 검증에 대한 loss가 가장 심한 격차를 보이며, CRNN은 학습 및 검증 데이터가 지속적인 감소를 보이는데 반해 한시적으로 검증에 대한 loss가 상승됨
- LFA-CRNN은 학습과 검증 데이터가 다소 격차를 보이는데 반해 심각한 격차를 보이지 않으며, 검증에 대한 loss가 한시적 증가가 발생되지 않음
- LFA-CRNN은 AlexNet 보다 안정적인 그래프를 그리며, 기존의 CRNN 보다 과적합 등과 같은 문제로 인해 그래프가 한시적으로 이탈 형상이 발생되지 않는 것으로 보아 LFA-CRNN은 비교 학습모델 보다 안정적이라 할 수 있음

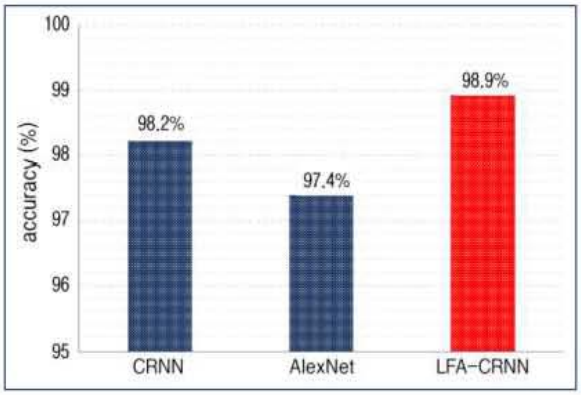
• UNBC-McMaster 데이터베이스를 이용한 손실 측정 결과



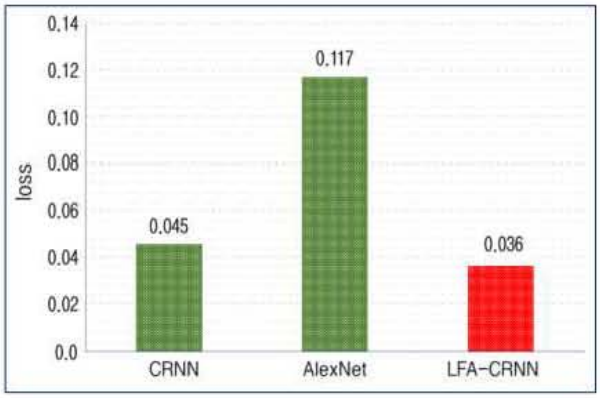
- LFA-CRNN이 약 98.92%로 가장 높은 정확도를 보이며, Loss 역시 약 0.036으로 가장 낮은 값을 보임
- CRNN은 학습 중 중간에 일시적으로 과적합이 발생되어 LFA-CRNN보다 낮은 정확도를 보인 것으로 예상

되고, AlexNet 역시 검증데이터의 심한 격차로 인하여 정확도 측정에서 성능이 저하된 것으로 판단함

• 테스트 데이터를 통해 정확도 및 loss 측정 결과



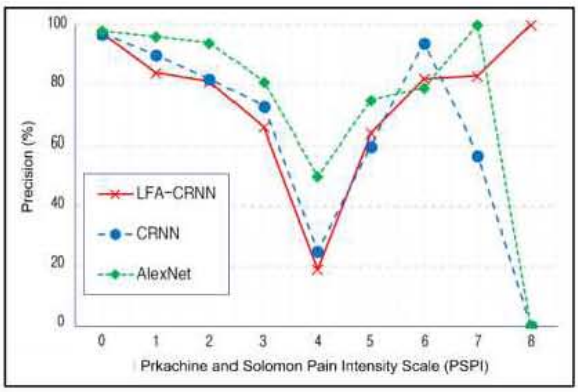
(a) Accuracy



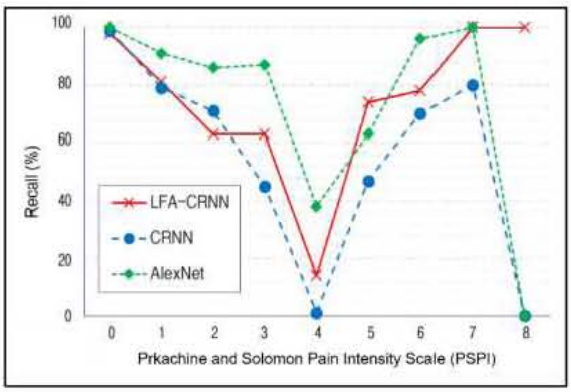
(b) Loss

- 재현율의 측정 결과는 각 고통 강도 클래스에 대하여 각 해당 클래스가 true라 예측한 샘플 중 실제도 true에 속하는 샘플 수의 비율을 나타냄
- LFA-CRNN은 "0"이 98%, "1" =81%, "2" =63%, "3" =63%, "4" =19%, "5" =74, "6" =78%, "7" =100%, "8" =100%로 AlexNet과 CRNN 보다 다소 저조한 결과를 보임. 차원 축소 LFA 기법으로 인하여 발생된 문제로 판단함. 차원 축소에서 원본 이미지를 압축하여 새로운 데이터를 생성 또는 강한 특징들로 구성된 일부의 특징만을 사용하여 데이터의 크기를 줄이는 것으로 세부 특징을 제거함으로써 강한 특징만을 사용하게 됨

• Precision 및 Recall에 대한 측정된 결과

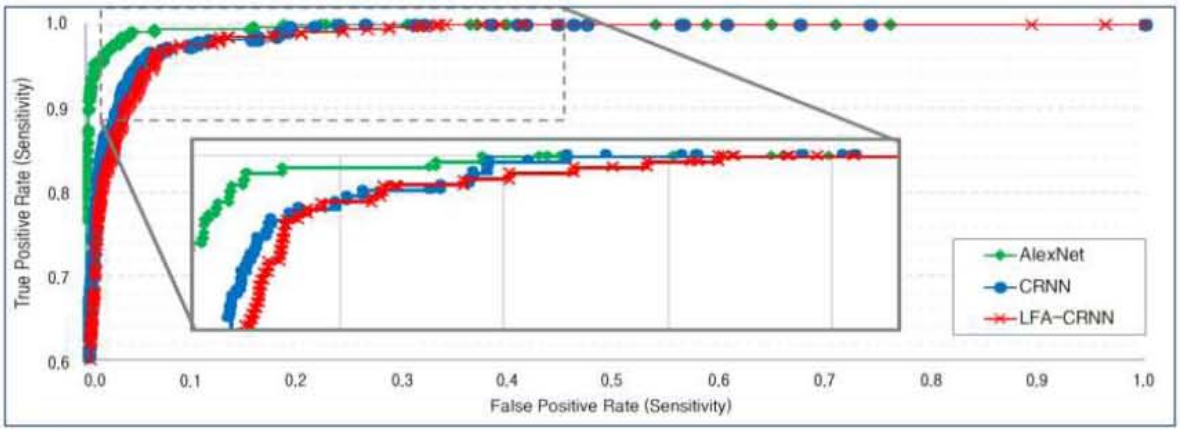


(a) Precision



(b) Recall

• ROC curve에 대한 측정된 결과



- PSPI가 “8” 인 데이터에서 LFA-CRNN 방법만 객체 검출이 가능하였음. 재현율을 평균적으로 확인할 경우 LFA-CRNN은 75%, AlexNet이 75%, CRNN이 0.56%로, LFA-CRNN은 AlexNet과 같이 75%를 결과를 나타냄
- 정확도는 재현율과 비슷한 결과를 보였으며, 평균적으로 LFA-CRNN은 75%로 가장 높으며, AlexNet이 73%, CRNN이 56% 순으로 측정됨
- 모든 모델이 PSPI가 “4” 강도를 보이는 데이터 검출하기 어렵다는 것을 보였으며, 강도 “8” 인 데이터에서 LFA-CRNN 방법만 객체 검출이 가능함을 보임
- 성능평가 결과, LFA-CRNN 모델은 학습 및 검증에서 안정적인 그래프를 보였으며, 테스트 데이터를 이용한 평가에서는 98.92%로 좋은 성능을 보임
- loss 역시 약 0.036으로 가장 낮은 결과를 보이며 정확도와 재현율의 결과에서 저조한 결과를 보였으나, 평균적으로 재현율은 75%로 AlexNet과 같이 높은 성능을 보였으며, 정확도는 75%로 높은 성능을 보임

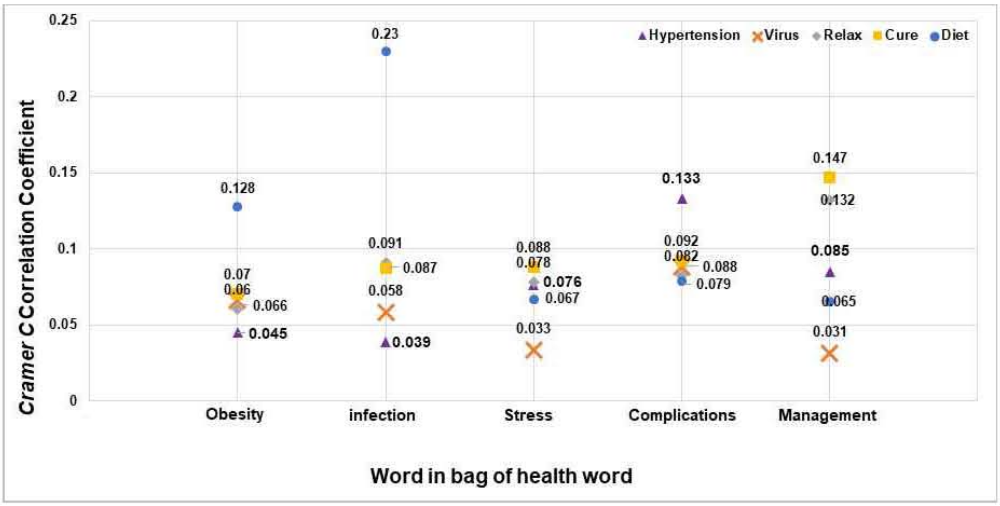
■ P2P 엣지 네트워크에서 다계층 헬스 지식 마이닝의 결과 및 성능평가[59)]

- 시간 복잡도는 알고리즘이 문제를 해결하기 위한 시간 또는 연산의 횟수를 의미하고 마이닝의 성능에 중요한 요소임. 빈발패턴탐색에서 빈발항목의 수를 a라 할 때 첫 번째 탐색 단계에서 a(a+1)/2개의 후보집합을 생성하고, 다음 단계에서도 후보항목을 생성하며 연관규칙 실행시간은 최소지지도 및 트랜잭션의 수에 따라 차이가 있음. 이는 최악의 경우 시간 복잡도는 $O(2^n)$임
- Apriori 연관규칙의 시간 복잡도에 $\log_2$를 적용하여 $\log_2 O(2^n)$로 계산하고, FP-Tree인 경우 m은 어떤 항목에 대한 빈발항목을 탐색하기 위해 포함하는 모든 경로의 합이 최소지지도를 만족하는 노드의 수를 나타내며, Apriori와 향상도 기반 상호정보량은 후보집합을 생성하지만 FP-Tree는 후보집합을 생성하지 않기 때문에 FP-Tree는 m의 값에 따라 시간복잡도가 지속적으로 증가함

• Apriori, FP-Tree를 이용한 연관규칙 알고리즘과 향상도 기반 상호정보량의 시간복잡도 비교 결과

	Apriori	FP-Tree	Ours
Number of Data Scan	C_n+1	2	C_n+1
Create Candidate Item	Yes	No	Yes
Time Complexity	$O(2^n)$	$O(n\times(2^m-1))$	$\log_2 O(2^n)$

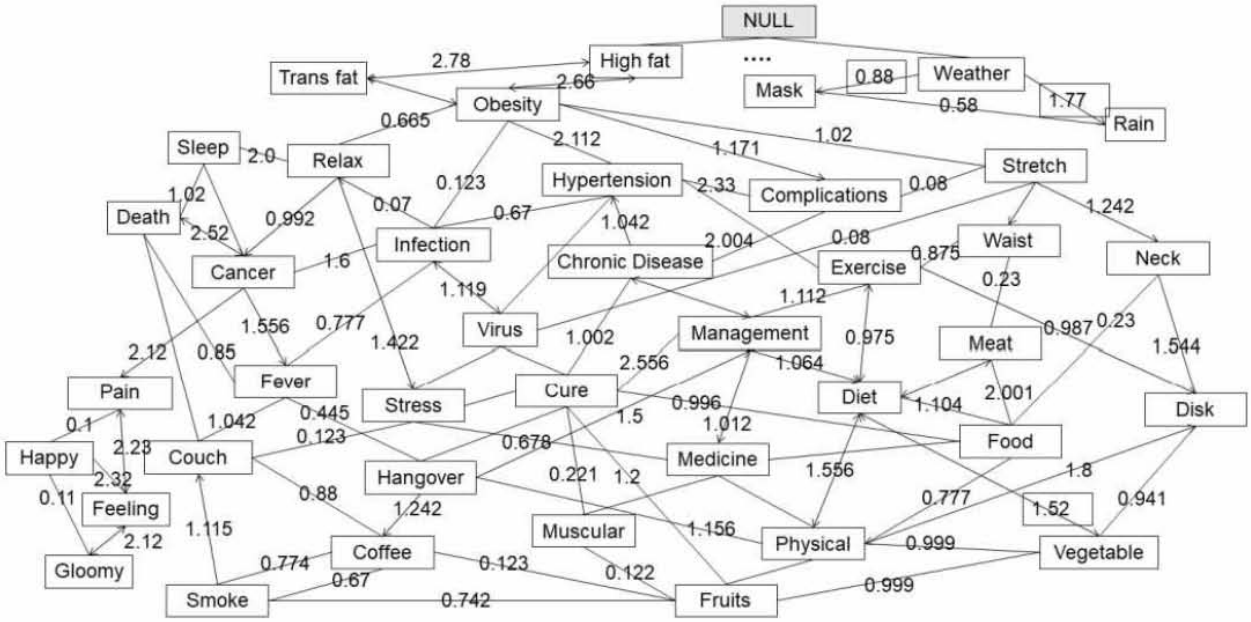
• 상호정보량 기반의 상관계수 그래프



59) J. W. Baek, K. Chung (2021) Multi-Level Health Knowledge Mining Process in P2P Edge Network. IEEE Access, 9, 61623-61634.

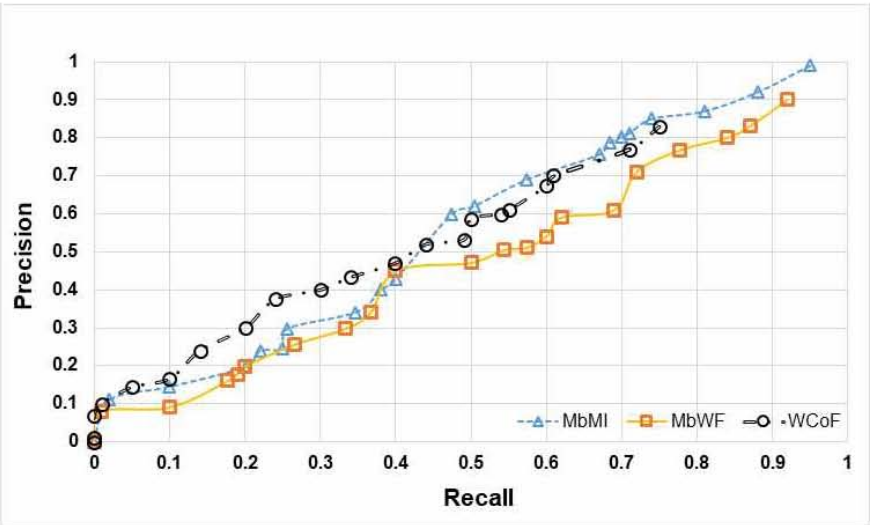
- Chronic Disease는 Hyperension와 상관계수가 0.045, Virus와 0.06, Cure와 0.07, Relax와 0.066, Diet와 0.128의 상관도를 나타냄. 상관계수가 0에 근접한 값이 추출되었기 때문에 상관도가 높음을 의미함
- IoT 장치를 사용하여 엣지 네트워크에 분산된 헬스 플랫폼에 헬스 데이터를 요청함. 엣지 네트워크는 분산된 헬스 서버에서 사용자가 요청한 정보를 실시간으로 처리 및 분석하고 메타정보를 생성하여 제공함. 메타정보는 텍스트 형태로 명사 추출과정을 통해 11,668개의 명사를 추출함
- 전처리 과정을 통해 10,400개의 단어가 추출하여 Bag of Health Word로 구성함. Bag of Health Word는 연관규칙 발견을 위한 트랜잭션으로 구성하고 Apriori 알고리즘과 Multi-Level Rule 알고리즘을 이용하여 연관관계를 발견함. 발견한 연관관계를 기반으로 상호정보량을 통해 연관 정도를 계산하고, 상관계수를 이용하여 상관 정도를 파악함

• 2P 엣지 네트워크에서 다계층 헬스지식 결과



- 단어들의 관계를 찾기 위해 제안하는 마이닝 기반 상호정보량(MbMI)과 기존에 존재하는 방법의 마이닝 기반 단어 빈도수(MbWF)와 문서 내의 단어 동시 출현 빈도수(WCoF)를 비교하기 위해 최소지지도를 반복적으로 변화시키며 성능평가를 진행

• MbMI, MbWF, WCoF에 따른 정확도와 재현율

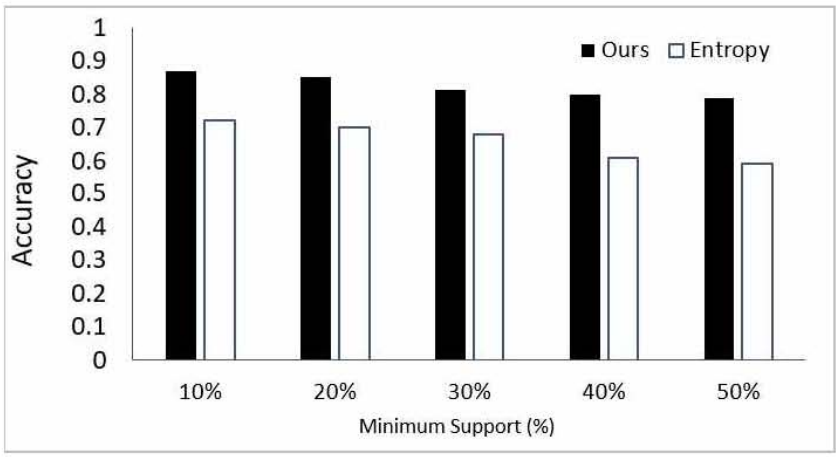


- MbMI 방법이 재현율을 기준으로 약 0.4부터 재현율과 정확도가 우수하게 평가됨. WCoF는 문서 내에서 단어간의 동시출현횟수가 제한되어 나타나므로 연관성을 고려하지 않아 재현율이 낮게 평가됨. MbWF는 단어

간의 상호 연관정도를 고려하지 않아 의미없는 관계가 발견되어, 정확도에서 낮게 평가됨

- 최소지지도의 변화에 따라 엔트로피를 이용한 정보량과 향상도 기반 상호정보량에 대한 정확도를 비교함. 향상도 기반 상호정보량이 엔트로피 기반 정보량에 비해 모든 지지도에서 높은 정확도를 보임

• 엔트로피와 향상도 기반 상호정보량의 정확도 비교 결과



- 향상도 기반 상호정보량의 우수성을 평가하기 위해서 기존의 존재하는 연관규칙 기반의 지식 마이닝 방법을 비교는 지지도 변화에 따라 규칙을 생성에 대해서 재현율, 정확도를 이용한 F-measure를 비교함

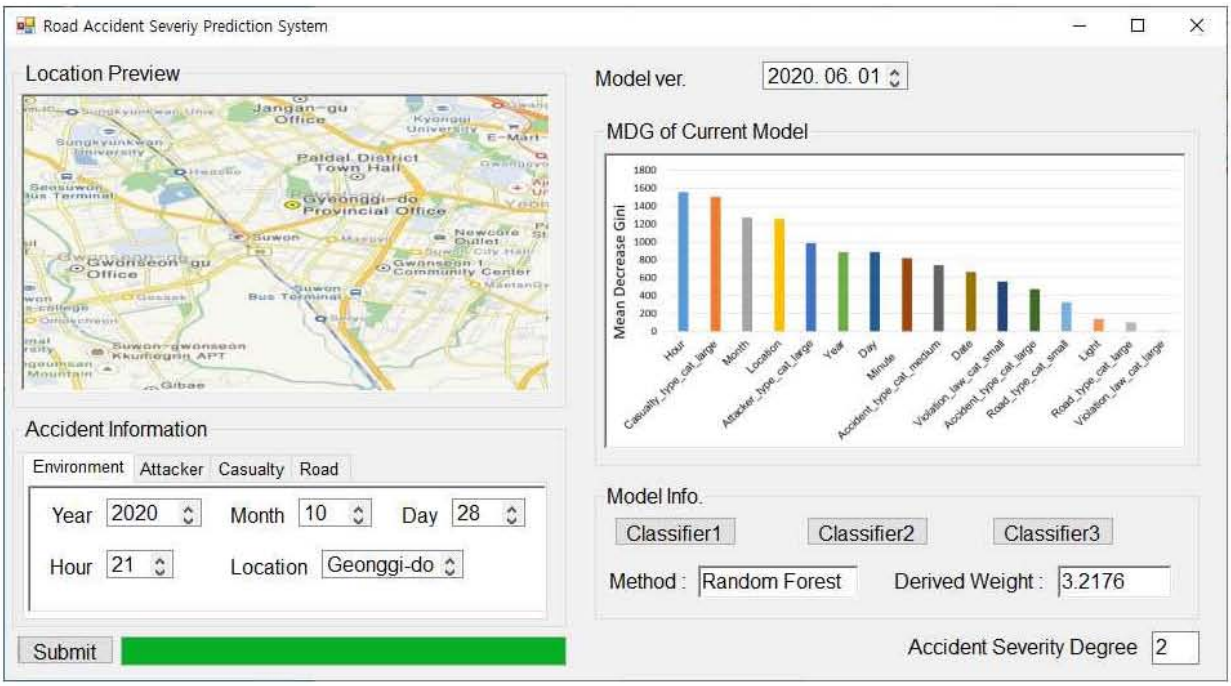
• 향상도 기반 상호정보량과 기존 모델과의 성능비교 결과

Method	Evaluation	10%	20%	30%	40%	50%
Ours	Precision	0.870	0.850	0.813	0.803	0.787
	Recall	0.810	0.740	0.711	0.700	0.684
	F-measure	0.839	0.791	0.759	0.748	0.732
R. Xu et al. [41]	Precision	0.735	0.705	0.613	0.683	0.658
	Recall	0.713	0.628	0.675	0.713	0.723
	F-measure	0.724	0.665	0.643	0.698	0.689
M. Nasr et al. [42]	Precision	0.682	0.653	0.513	0.588	0.495
	Recall	0.535	0.500	0.651	0.610	0.600
	F-measure	0.600	0.566	0.574	0.599	0.542

- 성능평가 결과 향상도 기반 상호정보량이 가장 우수하게 평가됨. R. Xu et al. 방법의 경우 상관관계 분석 및 비선형 모델링 방법을 통해 지식을 생성하며, 연관규칙 마이닝을 이용하여 위험 요인과 관리자의 관계를 추출하여 랜덤포레스트 알고리즘을 통해 지식을 추출을 위한 모델링을 하여 데이터가 위험한 상황만으로 제한되어 있기 때문에 성능이 낮게 평가됨
- M. Nasr et al.의 방법은 클래스 레이블이 지정된 데이터를 사용하는 지도학습 방법으로 지식을 추출하기 때문에 라벨이 없거나 새로운 데이터로 분석을 할 경우 재현율 및 정확도가 낮게 평가됨

■ 교통사고 심각도 예측을 위한 가중치 피드백 기반 Harmonic MDG-Ensemble 모델 평가[60]

• Harmonic MDG-ensemble 모델을 이용한 교통사고 심각도 예측 시스템



- 사용자는 해당 요소에 대해 전처리를 진행하지 않고 설문 조사와 유사하게 시스템의 각 항의 해당하는 부분에 표기함. 사용자의 데이터 입력이 종료되면 시스템은 구축된 교통사고 심각도 예측 모델에 해당 데이터를 투입하며 입력값에 해당하는 교통사고 심각도 범주 값이 결과로 도출됨
- 앙상블 모델과 단일 알고리즘의 랜덤포레스트, KNN, XgbTree의 분류 모델에 따른 성능 비교함

• 성능평가 비교 결과

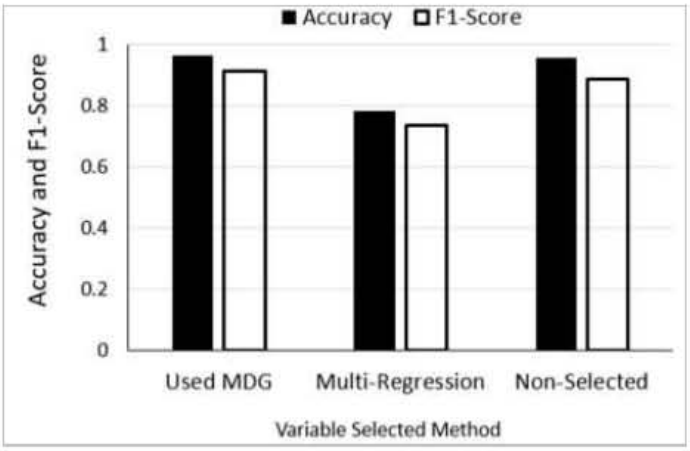
	RandomForest	KNN (K-Nearest Neighbor)	xgbTree	Ours
Accuracy	0.9655	0.8364	0.9553	0.9686
Recall	0.8533	0.8738	0.9743	0.9446
Precision	0.9830	0.4462	0.8148	0.9089
F1-Score	0.9136	0.5907	0.8875	0.9264

- 앙상블 모델이 단일 모델보다 Accuracy와 F1-score가 우수하게 평가됨. recall, precision의 조화평균인 F1-score 면에서는 제안하는 앙상블 모델이 가장 높은 수치를 가지므로 제안하는 Harmonic MDG-ensemble 모델이 이를 구성하는 단일 모델보다 일반적인 성능이 뛰어난 것을 의미함

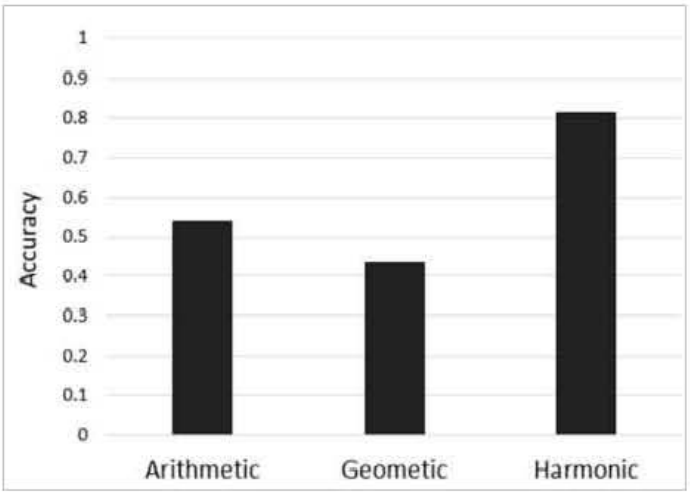
- 종속변수에 대한 영향력이 큰 독립변수만 채택하는 변수선택 과정 방법에 따른 성능을 비교함. 변수선택 과정에 대한 비교는 제안하는 MDG와 다중회귀분석 방법, 변수선택 과정 없을 경우를 비교하여 정확도와 F1-Score를 비교함
- 종속변수를 교통사고 심각도로 하며, 독립변수는 16개의 변수를 사용함. 변수 선택 후 MDG 모델은 11개의 독립변수를 사용하며 다중 회귀방법은 유의수준 0.5 이하를 만족하는 10개의 독립변수를 사용함

60) B. K. Koo, J. W Baek, K. Chung (2021) Weight Feedback based Harmonic MDG-ensemble Model for Prediction of Traffic Accident Severity, Applied Sciences, 11(11), 5072-5089.

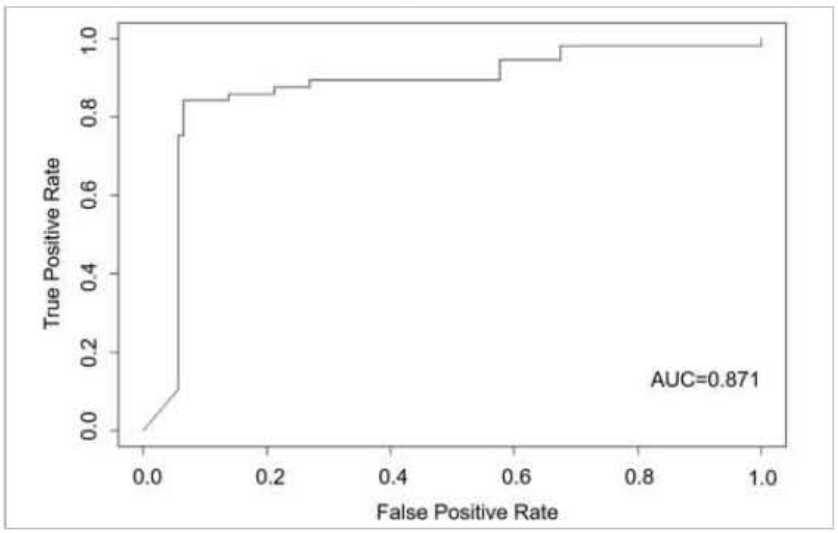
• 평균 감소 지니계수 기반 모델과 비 지니 모델 간의 성능 비교 결과



• 산술평균, 기하평균 및 조화평균 간의 정확도 비교 결과



• 모델의 ROC(Receiver Operating Characteristic) 및 AUC(Area Under Curve) 영역



- 산술평균을 적용할 경우 MDG 계수를 통해 추출한 시간, 사상자수, 월, 위치, 가해자 유형, 년도, 분, 사고유형, 날짜, 법률위반, 요일의 11개 변수가 공통적인 기준이 없기 때문에 나눌 수 있는 기준이 없으므로 산술평균을 이용했을 경우 정확도가 낮게 평가됨
- 11개의 변수가 일정하게 증가하는 비율을 찾을 수 없으므로 기하평균을 적용할 경우 정확도가 가장 낮게 평가됨. 제안하는 조화평균을 적용할 경우 균형을 유지하며 변화하는 평균 비율을 판단하므로 성능이 가장 우수하게 평가됨

• 성능 비교 결과

	Accuracy	Recall	Precision	F1-Score
Proposed	0.9686	0.9446	0.9089	0.9264
Hashmienejad et al. [39]	0.8820	0.8345	0.8856	0.8875
Sharma et al. [6]	0.8606	0.7630	0.8114	0.7865

- 제안하는 방법의 성능이 Accuracy, Recall, Precision, F1-Score에서 Hashmienejad, Sharma 방법 대비 우수한 성능을 가짐. Hashmienejad의 모델은 사용자의 선호도에 따라 교통사고 심각도를 분석하였으나 전문가의 선호도를 기반으로 규칙을 생성함
- 전문가에 따른 각각의 데이터 셋에 따라 결과가 달라지기 때문에 제안하는 방법에 비해 성능이 낮게 평가됨. Sharma의 방법은 분석하는 데이터의 특성과 맞지 않는 문제점이 있어 성능이 낮게 평가됨

■ 지식 기반 도심지 교통사고 위험도 예측 모바일 서비스 개발[61]

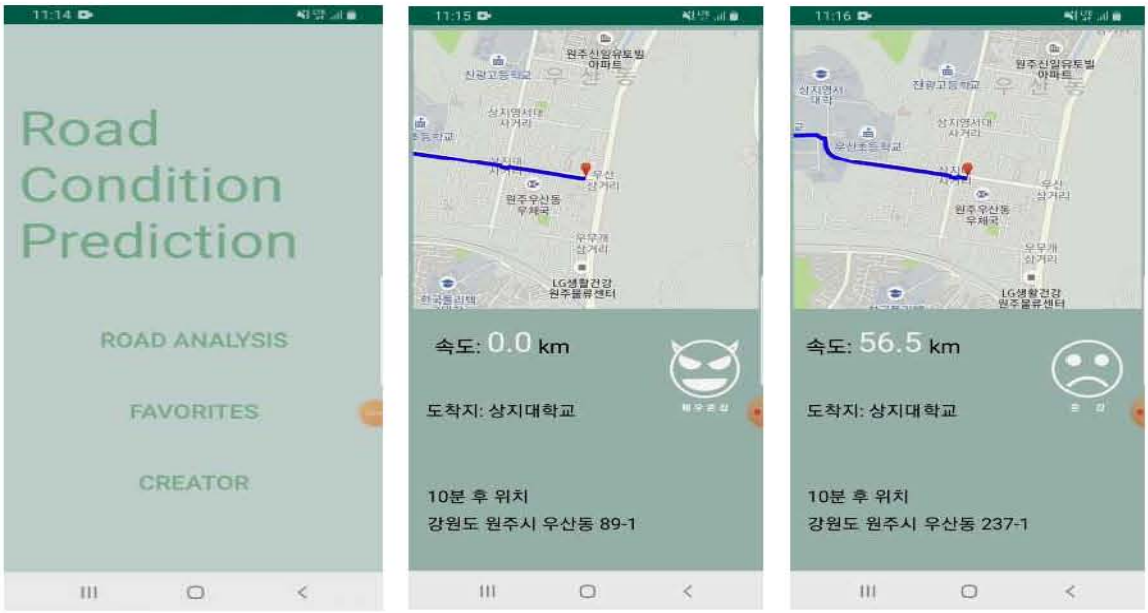
- 공공 사고 데이터를 기반으로 도로의 정적인 특성(도로의 길이, 제한속도 등)과 동적인 특성(시간별 교통량, 태양의 고도, 방위각, 건물의 고도 등)의 특성을 반영하여 실시간 도로 위험도 예측 및 안내함
- 딥러닝에 기반한 도로 사고위험도를 예측하고 예측된 결과를 반영하여 기존 통계 기반의 단순 사고다발지역 안내가 아닌 같은 도로라도 변화하는 상황에 맞는 위험도를 안내한다. 사고 발생에 영향을 미치는 요인으로 도로의 고정된 특성인 정적 특징과 시간과 상황에 따라 변하는 동적 특징을 설정하고 이와 관련된 사고위험도를 예측한다. 사용자의 경로에 따른 도로소통 상황 및 교통사고 다발 지역, 실시간 도로 통제 상황 등을 고려한 사고위험도를 제공함

- 딥러닝에 기반한 도로 사고위험도를 예측하고 예측된 결과를 반영하여 기존 통계 기반의 단순 사고다발지역 안내가 아닌 같은 도로라도 변화하는 상황에 맞는 위험도 안내
- 사고 발생에 영향을 미치는 요인으로 도로의 고정된 특성인 정적 특징과 시간과 상황에 따라 변하는 동적 특징을 설정하고 이와 관련된 사고위험도를 예측함
- 사용자의 경로에 따른 도로소통 상황 및 교통사고 다발 지역, 실시간 도로 통제 상황 등을 고려한 사고위험도 제공

- 교통사고와 관련된 다양한 인적, 도로 기하구조 특성과 환경적 요인들과 교통사고 간의 인과관계를 찾고 분석하는 통계적 분석과 자료 분포 및 함수 형태에 대한 가정에 따른 한계를 해결하였음
- 딥러닝에 기반한 사고위험도 예측 모델을 개발하여 도로의 길이, 제한속도 등과 도로의 정적인 특성과 주행하는 시간에 따른 교통량, 태양의 고도와 방위각과 같은 동적인 특성을 반영하여 도로 위험도를 예측하는 모델을 개발하였음
- 2020년 8월~12월의 5개월 동안 경찰청 도시정보교통센터 및 ITS 국가교통정보센터에 의해 수집된 데이터 4,470건을 활용하여 도심내 단속류에 대해 모델을 구축하고 테스트한 결과, 정확도는 75%, 재현율은 81%로 예측 모델의 성능을 확인하였음
- 개발된 위험도 예측 모델의 결과에 기반하여 기존 통계 기반의 단순 사고다발지역 안내가 아닌 같은 도로라도 변화하는 상황에 맞는 위험도를 안내하는 시스템을 구축하였음. 안내 시스템은 도로구간 또는 지역별 위험도 정보제공뿐만 아니라 안전시설 설치 등과 같은 도로 개선에 활용하였고 운전자에게 안전을 위해 최적화된 경로를 제공하는 모바일 시스템을 구축하여 서비스를 제공함
- 국토지리원에서 제공하는 지리 정보와 ITS 교통정보센터에서 제공하는 표준 노드링크 정보를 정합하여 도로에 인접한 5층 이상의 상가건물로 간주하여 건물 수 정보로 활용하였음. 단순 시간대 정보가 아닌 차량 운행 시간과 주행 방향에 따라 달라지는 태양 고도와 방위각을 학습 특징으로 활용하여 운전자 시야에 끼치는 영향을 학습하는데 활용하였음

61) 지식기반 도심지 교통사고 위험도 예측 모바일 서비스, 등록번호 C-2021-041263, 한국저작권위원회, 2021.10.19.

• 지식기반 도심지 교통사고 위험도 예측 서비스 화면



- 특징별 사고발생 빈도 간의 관계성 분석을 통해 사고와의 관련성을 분석하였고, 도로에 인접한 건물의 수가 많을수록 사고 발생 빈도가 더 높고, 야간보다 통행량이 많은 주간에 사고 발생 빈도가 더 높음을 알수 있음. 또한, 태양을 마주보고 주행할수록 사고빈도가 더 높은 의미있는 결과를 도출하여 해당 패턴 특징이 모두 사고와 관련성이 있음을 확인하였음

■ 빅데이터 기반 교통안전 이머징 리스크 마이닝 어플리케이션 개발[62]

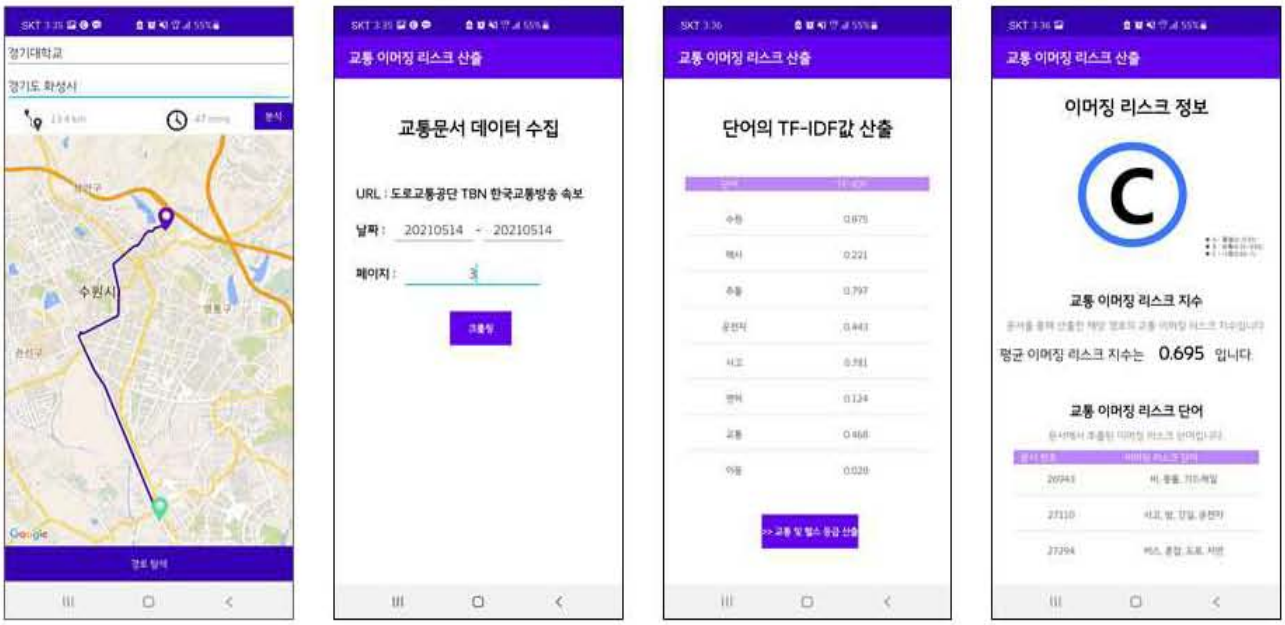
- 교통 사고율이 증가함에 따라 교통위험 탐지에 대한 중요성이 증가하고 있음. 사용자의 위치와 경로에 따라 적절한 교통정보를 제공할 필요성이 있기 때문에 이를 고려한 분석 방법이 필요함. 교통정보는 예상하지 못한 새로운 이벤트들이 발생되며, 정보의 유효기간성이 짧기 때문에 과거의 데이터만을 이용해 교통상황을 예측하는 것이 어려움. 따라서 과거의 데이터뿐만 아니라 실시간적인 데이터를 활용하여 운전자에게 교통정보를 쉽고 명확하게 전달할 수 있도록 교통 데이터와 운전자의 위치정보 등을 고려한 클라이언트 기반의 모바일 인터페이스 개발함
- 비정형 데이터에 대한 분석의 가치의 중요성이 높아지고 있고, 비정형 데이터의 형태가 정해져 있지 않기 때문에 값의 의미를 쉽게 파악하기 어려움이 존재함. 각 데이터의 특성에 따른 마이닝을 통해 유의미한 정보를 추출하는 방법이 필요함[63]. 비정형 데이터 분석을 위해 텍스트 마이닝에 대한 결과를 적용함
- 텍스트 마이닝은 전처리 된 텍스트에서 연관규칙, 군집분석, 분류 등의 알고리즘을 이용하여 정보를 추출할 수 있는 기술임. 데이터의 지속적인 증가로 인해 대량의 텍스트 데이터에서 리스크에 관련된 정보를 추출하는 것에 대한 어려움이 존재함. 텍스트에 내재된 의미를 파악하여 리스크에 대한 정보를 파악이 가능함
- 텍스트를 통해 오피니언 마이닝을 이용한 감성분석이 가능함. 감성분석은 텍스트가 지니고 있는 감성을 분석하기 위해 수치화된 객관적인 정보로 변환하여 긍정, 부정, 중립에 대한 의견을 분류하여 의사결정에 적용함. 단어에는 동음이의어가 존재하기 때문에, 단어를 같게 읽지만 뜻은 다르기 때문에 키워드를 통해 위험 정보를 예측하기에는 혼란이 생김. 따라서 리스크 스코어 및 등급을 통해 시각화하여 위험도에 대한 정보를 제공함
- 텍스트 데이터에는 다양한 정보들이 내재되어 있지만 텍스트만을 이용하여 리스크를 인지하고, 예측하는 것은 어려움이 존재하기 때문에 텍스트 문서를 수집하고 형태소 분석을 통해 키워드를 추출하여 단어의 중요도를 파악함. 키워드 간의 감정 유사도 가중치를 통해 기존의 감성사전에 존재하지 않은 단어도 분류 및 예측이 가능하도록 새로운 사전을 구축하여 제한된 단어 사용 문제를 해결함
- 사용자의 위치와 출발지 및 도착지 입력을 통해 경로를 찾을 수 있으며, 사용자가 직접 입력한 출발지부터

62) 빅데이터 기반 교통안전 이머징 리스크 마이닝 어플리케이션, 등록번호 C-2021-024308, 한국저작권위원회, 2021.06.17.
63) M. J. Kim, J. S. Kang, K. Chung (2020) Word-Embedding-Based Traffic Document Classification Model for Detecting Emerging Risks Using Sentiment Similarity Weight, IEEE Access, 8, 183983-183994.

도착지까지의 거리와 시간을 확인할 수 있음. 사용자가 도로교통공단 TBN 한국교통방송[64] 속보의 날짜와 웹페이지에서 수집할 페이지를 입력하고 크롤링을 통해 데이터를 수집함. 수집된 교통문서에서 키워드를 추출하기 위해 형태소분석을 진행하여 추출한 키워드의 중요도를 파악하기 위해 TF-IDF 값 산출

- 모든 경로에 따른 위험지수를 계산하여 사용자의 인터페이스에 위험도를 시각화하는 것은 상당한 비용과 시간이 소모되기 때문에 사용자의 최적의 경로를 파악하고, 해당 경로에서 발생된 교통이벤트만을 감지하여 운전자의 위치를 기반으로 이머징 리스크에 대한 교통안전지수를 산출함

• 빅데이터 기반 교통안전 이머징 리스크 마이닝 어플리케이션 화면



- 어플리케이션에서 제공하는 이머징 리스크는 텍스트 데이터에서 유의미한 정보를 추출하여 제공함. 이는 기본의 거리 기반 유사도 방식은 단어의 빈도만 고려하기 때문에 같은 맥락의 분서들을 다르게 판단하는 분제점이 있음. 단어의 방향성을 고려하는 코사인 유사도를 통해 각 키워드 간의 유사도를 추출하여 감성 사전을 재구성함. 이는 기존의 감정 사전에 존재하지 않은 단어도 사용 가능한 장점이 있으며, 다양한 텍스트 분서 간의 유사도 추출이 가능함
- 이머징 리스크 등급은 사용자의 위치 경로와 관련된 정보를 제공하며, 사용자가 방대한 양의 텍스트를 분석하지 않고, 시각화된 정보를 통해 리스크를 판단할 수 있음

■ 이머징 리스크 예측을 위한 교통안전 지수 프로그램 개발[65]

- 사용자의 경로에 따른 도로소통 상황 및 교통사고 다발 지역, 실시간 도로 통제 상황 등의 종합 교통정보 제공하고, 기상청에서 실시간으로 제공하는 사용자의 안전과 관련된 보건기상지수, 생활기상지수, 대기오염지수를 제공하여 헬스 리스크 정보 안내함
- 사용자 경로 기반 교통위험 요소와 실시간 헬스리스크 정보를 종합한 이머징 리스크 안전지수 제공하고, 과거의 사고 발생 데이터를 기반으로 예상 경로에서의 이머징 리스크 발생 고위험 지역 정보 안내함
- 실시간으로 사용자의 안전과 관련된 보건 기상지수, 생활기상지수, 대기오염지수를 제공하여 헬스 리스크 정보 안내하고, 휴먼 인터페이스 소프트웨어 개발 및 사고위험도 예측을 위한 실시간 멀티미디어 서비스를 제공. 정확한 위험도 예측 서비스를 위해 시공간 결측 보정을 통한 잠재적 교통사고 위험도 분석, 이머징 리스크 마이닝 인터렉티브 휴먼 인터페이스를 제공

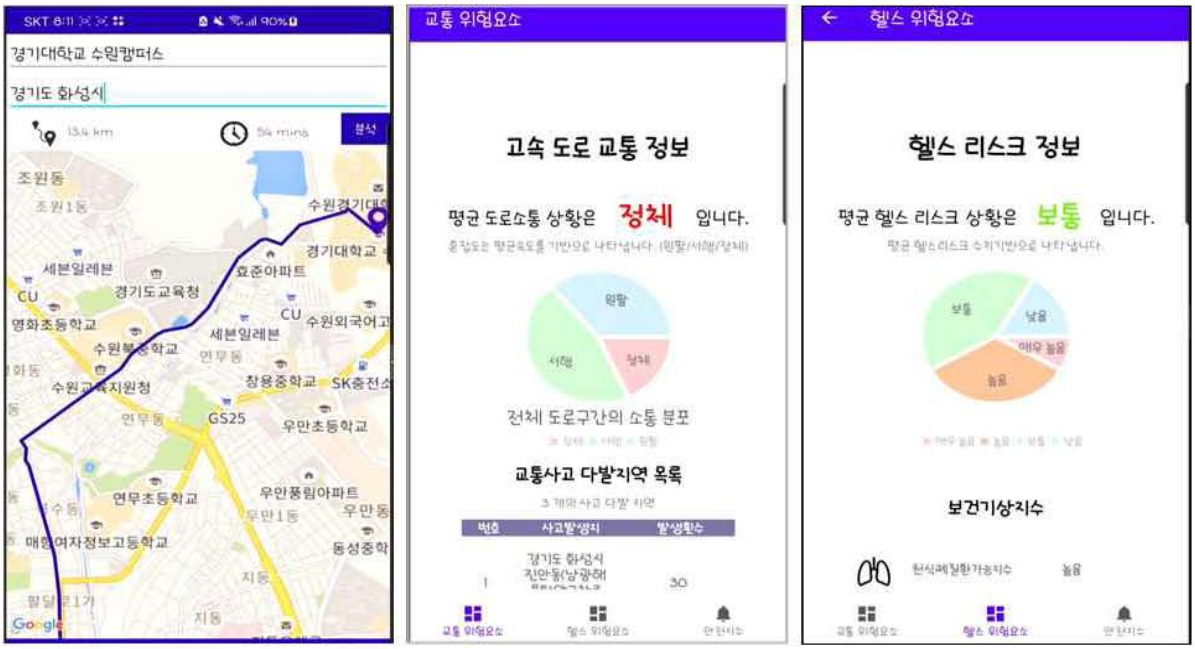
- 사고위험도 예측 경로에 기반하여 가장 안전하고 빠른 경로 추천 및 정보 시각화가 가능하며, 구간에 대한 교통상황 예측정보 및 시간별 도로위험도 점수 산출하여 제공함. 도로교통 이용자는 안전 지수를 통해 정

64) Traffic Broadcasting Network, https://www.tbn.or.kr/.
65) 이머징 리스크 예측을 위한 교통안전 지수 프로그램, 등록번호 C-2021-024309, 한국저작권위원회, 2021.06.17.

보를 쉽고 빠르게 인지하여 교통안전 지수에 대한 리스크 상황을 제공함

- OPEN API와 XML 형식의 교통 빅데이터 기반으로 다양한 교통요인과 운전자의 교통정보를 고려하여 경로와 경로의 주행거리 및 시간을 제공 및 이머징 리스크에 대한 교통안전 요인을 분석하여 제공함

• 이머징 리스크 예측을 위한 교통안전 지수 프로그램 화면



■ 영상 콘텐츠의 연관지식 그래프 생성 및 결과[66]

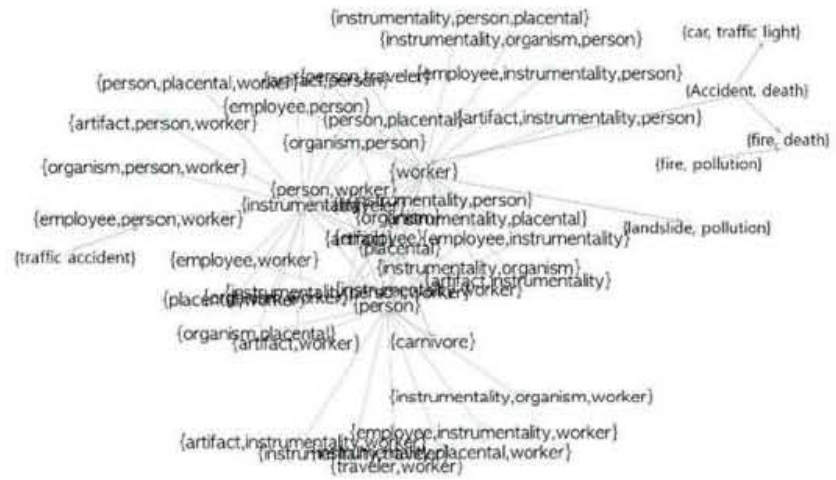
- 영상 콘텐츠의 연관지식 그래프 생성을 위해 제안하는 방법을 통해 생성한 Fuzzy C-Means 군집을 이용하여 각 군집별로 분류된 영상을 통해 트랜잭션을 재구성하여 연관규칙 알고리즘을 적용하여 연관지식 그래프를 추출함
- Fuzzy C-Means 군집을 이용하여 클러스터 1로 분류된 영상은 총 362개, 클러스터 2는 308개의 영상 콘텐츠가 분류됨. 각 클러스터로 분류된 영상의 총 개수가 다르기 때문에 데이터 불균형의 문제 발생함. 실제로 하나의 영상 정보를 나타내는 트랜잭션 열의 양이 적으면 적을수록 높은 지지도가 발생함
- 만약 670개의 영상 데이터 중 하나의 영상이 특정한 A => B에 해당하는 아이템셋을 가지고 있을 확률은 1/362이지만, 군집 2의 경우 1/308로 그 값의 비중이 크기 때문에 같은 지지도 값을 바탕으로 Pruning을 진행하였을 때, 오히려 적은 양의 정보를 가지고 있는 군집을 생성할 수 있는 연관규칙의 수가 더 많은 단점이 존재함

- 군집 중심과의 거리를 기준으로 데이터를 컷오프 하여 최종적으로 각 클러스터당 70개의 영상 데이터로 제한하여 각 영상이 연관규칙 알고리즘에 미치는 영향의 정도를 동일하게 하여 정확한 지식 그래프의 생성을 진행함
- 군집 1에서 362개의 영상을 선택하기 위해 각 군집의 중심점과 가장 가까운 순서로 정렬하고, 군집의 중심점과 가장 가까운 70개의 영상을 이용하여 트랜잭션을 재구성함. 재구성한 각 군집별 트랜잭션을 이용하여 영상 지식 그래프 생성하여 자연 및 생활환경과 인간의 문화에서 연관되어 발생할 수 있는 이머징 리스크의 관계를 파악함
- 지식 그래프의 생성을 위해 Apriori 기반 연관규칙을 이용하여 연관규칙의 결과물을 그래프로 표현함. 연관규칙은 아이템간의 연관성을 파악하기 위해 단어의 관계를 지지도와 신뢰도를 기준으로 파악함. 재구성한 트랜잭션을 바탕으로 불필요한 규칙을 제거하고 의미있는 규칙을 추출하기 위해 최소지지도 0.6, 최소 신뢰도 0.5로 pruning을 진행한 후, 연관규칙 알고리즘 적용함
- 지식 그래프는 Lhs => Rhs 규칙을 이용하여 edge와 vertex에 적용함. 지식 그래프는 방향성이 존재하는 유

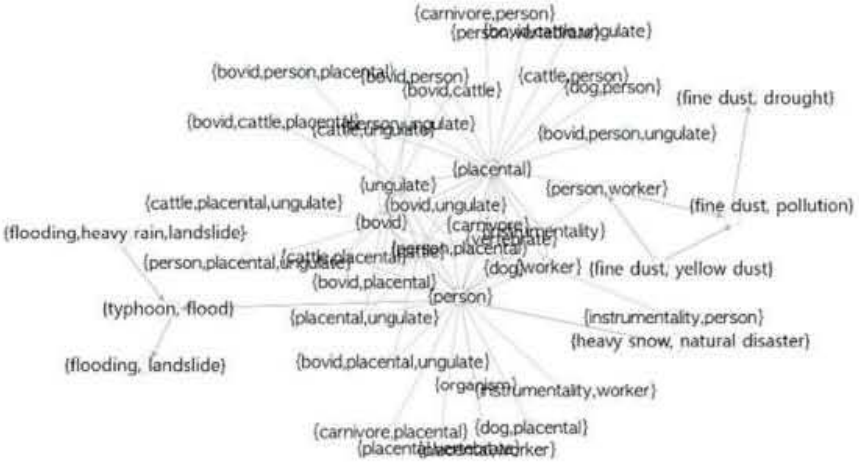
66) H. J. Kim, J. W. Baek, K. Chung (2021) Associative Knowledge Graph Using Fuzzy Clustering and Min-Max Normalization in Video Contents, IEEE Access, 9, 74802-74816.

향그래프의 형태로 출력되며, 노드는 영상 콘텐츠에서 추출된 객체 데이터의 아이템 셋을 나타냄. SR Min-Max 정규화 과정과 연관 퍼지 C-평균 군집을 통해 재구성한 트랜잭션을 이용하여 지식 그래프를 생성

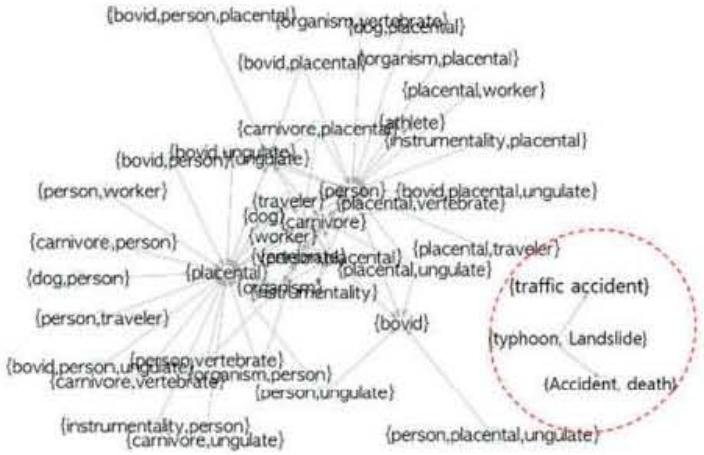
• Fuzzy C-Means 군집을 이용하여 연관지식 그래프 결과



(a)



(b)



- 클러스터링(a)의 지식 그래프에서는 traffic accident, fire, pollution 등의 교통, 환경오염으로 사람들에 의해 발생하는 인적재난에 주로 사용되는 규칙 발견함. 클러스터링(b)의 지식 그래프에서는 pollution, fine dust, yellow dust 등의 환경오염, 태풍, 홍수 등으로 자연재난에 주로 사용되는 규칙 발견함
- 인적재난은 화재, 산불, 사고 등의 사람에 의해 발생하는 재난을 의미하고, 자연 재난은 홍수, 태풍, 가뭄, 황사, 미세먼지 등의 자연현상으로 인해 발생함. 인적재난과 자연재난은 공통적으로 동물의 자연 및 생활환

경뿐만 아니라 인간에게도 잠재적인 위험 발생. 클러스터링(c)의 지식 그래프에서는 인적재난과 자연재난에서 동시에 나타나는 규칙으로 지식 그래프를 생성함

(2) 정량적 연구개발성과

○ (A-1)**위험 예측을 위한 다중회귀기반 상황 심층 신경망 모델 개발**
- 멀티-모달 Denoising 오토인코더 모델, 결측치 추론을 통한 정확도 향상
- 협력적 레이어 기반 하이브리드 멀티-모달 딥러닝
- 연구결과를 **SCI논문 2편 게재**
  - IEEE Access 2020 8 104933-104943 게재
    Multi-Modal Stacked Denoising Autoencoder for Handling Missing Data in Healthcare Big Data
  - IEEE Access 2020 8 192469-192480 게재
    Hybrid Multi-modal Deep Learning using Collaborative Concat Layer in Health Bigdata

○ (A-2)**상황 딥러닝 모델을 통한 이머징 리스크 상황 예측**
- 상황 딥러닝을 통한 이머징 리스크 예측
- 의사결정 경계 기반 이상치 탐지 모델, 교통영상의 Yolo 기반 광원 객체탐지
- 연구결과를 **국내등재논문지(비SCI) 1편 및 SCI논문 1편 게재**
  - 융합정보논문지 2020 10(8) 40-46 게재
    교통 영상 빅데이터 처리를 위한 Yolo 기반 광원 객체 탐지
  - IEEE Access 2020 8(1) 108664-108674 게재
    Decision Boundary-Based Anomaly Detection Model using Improved AnoGAN From ECG Data

○ (A-3)**잠재적 위험에 따른 사용자별 교통안전지수 개발**
- 순위스코어 연관지식 그래프 최적화
- 사용자별 잠재적 위험에 따른 교통안전지수 도출
- 연구결과를 **SCI논문 2편 게재**
  - Applied Sciences 2020 10(13) 4590-4610 게재
    Optimization of Associative Knowledge Graph using TF-IDF based Ranking Score
  - IEEE Access 2020 8(1) 183983-183994 게재
    Word Embedding based Traffic Document Classification Model for Detecting Emerging Risk using Sentiment Similarity Weight

○ (B-1)**사용자별 라이프로그 기반 헬스정보 분석 모델 개발**
- 멀티스레딩을 통한 빅데이터의 병렬 수집 모듈 개발
- 멀티 레벨 헬스 지식 마이닝 프로세스 개발
- 연구결과를 **SCI논문 1편 게재**
  - IEEE Access 2021 9(1) 61623-61634 게재
    Multi-Level Health Knowledge Mining Process in P2P Edge Network

○ (B-2)**시공간 결측 보정을 통한 LSTM 기반 교통위험 예측 모델 개발**
- LSTM 딥러닝 모델 정확도 성능 고도화
- LSTM 기반 리스크 예측 모델, 교통사고 심각도 예측 고도화
- 교통사고 심각도에 영향을 주는 변수 예측
- 연구결과를 **국내등재 학술지(비SCI) 1편 게재 및 특허 1건 출원**
  - 한국융합학회논문지 2021 12(2) 1-6 게재
    도로 CCTV 데이터를 활용한 딥러닝 기반 차량 이상 감지
  - 특허출원, 출원번호 10-2021-0078113, 2021.06.16

LSTM 기반 교통 혼잡도 예측 방법 및 장치

○ (B-3)**빅데이터 기반 교통안전 이머징 리스크 마이닝 예측**
- 데이터 편향성을 고려한 연관 마이닝 기반 도로위험 탐지 모델
- 숨겨진 연관정보에서 마이닝 기반 지식 그래프 발견
- 이상 상황에 대한 안전도 예측 모듈 개발
- MOG2를 이용한 배경 추출 및 딥러닝 기반 차량 이상 감지 알고리즘 개발
- 연구결과를 **SCI논문 3편 및 국내등재 학술지(비SCI) 1편 게재, SW등록 2건, 특허(출원) 1건**
  - IEEE Access 2021 9(1) 74802-74816 게재
    Associative Knowledge Graph using Fuzzy Clustering and Min-Max Normalization in Video Contents
  - Applied Sciences 2021 11(11) 5072-5089 게재
    Weight Feedback based Harmonic MDG-ensemble Model for Prediction of Traffic Accident Severity
  - Multimedia Tools and Applications 2021 80(26) 34609-34624 게재
    Multimedia Knowledge-based Bridge Health Monitoring using Digital Twin
  - 융합정보논문지 2021 11(7) 14-20 게재
    도로교통 이머징 리스크 탐지를 위한 AutoML과 CNN 기반 소프트 보팅 앙상블 분류 모델
  - SW저작권등록, 등록번호 C-2021-024308, 2021.06.17.
    빅데이터 기반 교통안전 이머징 리스크 마이닝 어플리케이션
  - SW저작권등록, 등록번호 C-2021-024309, 2021.06.17.
    이머징 리스크 예측을 위한 교통안전 지수 프로그램
  - 특허출원, 출원번호 10-2021-0046969, 2021.04.12.
    워드임베딩을 이용한 감성 유사도 가중치 기반 교통위험 탐지 및 예측 모델

○ (C-1)**교통혼잡도를 이용한 사고위험도 예측모형 개발**
- LSTM 교통혼잡도 예측 및 정확도 향상
- 시계열 특성과 상황별 학습 및 예측
- 연구결과를 **SCI논문 1편 게재**
  - IEEE Access 2020 8(1) 150784-150796 게재
    Prediction of Traffic Congestion Based on LSTM through Correction of Missing Temporal and Spatial Data

○ (C-2)**운전자 안면인식 기반 사고위험도 예측모형 개발**
- 차원축소 및 운전자 안면인식 상태 판단, 도로교통 이머징 헬스 리스크 발견
- 데이터 편향성 최적화 연관 추론 모델
- 연구결과를 **국내등재논문지(비SCI) 1편 및 SCI논문 1편 게재**
  - 한국융합학회논문지 2020 11(9) 1-6 게재
    도로 위험 탐지를 위한 데이터 편향성 최적화 기반 연관 추론 모델
  - Applied Sciences 2020 10(8) 2956-2975 게재
    Driver Facial Expression Analysis Using LFA-CRNN-Based Feature Extraction for Health-Risk Decisions

○ (D-1)**딥러닝 기반 미래 교통 혼잡도 예측 시스템의 고도화**
- 클라우드 기반 소셜 네트워크 구성 매커니즘 개발
- 프레임 전송간의 유휴시간(idle time) 감소
- 딥러닝 기반 교통 혼잡도 예측 시스템 성능 향상
- 연구결과를 **SCI논문 1편 게재 및 특허(출원) 1건**
  - Applied Sciences 2021 11(18) 8621-8640 게재
    Health Risk Detection and Classification Model using Multi-Model-based Image Channel Expansion

and Visual Pattern Standardization

- 특허출원, 출원번호 10-2021-0109154, 2021.08.19
  특징 추출 기반의 LFA-CRNN을 이용한 운전자 표정 분석 장치 및 그 방법

○ (D-2)**사고 위험도 예측 모바일 서비스 개발**

- 가장 안전한 경로 추천 및 정보 시각화 모듈 개발
- 정체 예측정보 및 시간별 도로 위험도 점수 제공 모듈 개발
- 시각화된 정보를 통한 사용자의 이해 시간 감소
- 연구결과를 **SW등록 1건**
- SW저작권등록 등록번호 C-2021-041263, 2021.10.19
  지식기반 도심지 교통사고 위험도 예측 모바일 서비스

(3) 세부 정량적 연구개발성과

[과학적 성과]

□ 논문(국내외 전문 학술지) 게재

번호	논문명	학술지명	주저자명	호	국명	발행기관	SCIE 여부 (SCIE/비SCIE)	게재일	등록번호 (ISSN)	기여율
1	Multi-Modal Stacked Denoising Autoencoder for Handling Missing Data in Healthcare Big Data	IEEE ACCESS	정경용	8	미국	IEEE	SCI	25 May 2020	10.1109/ACCESS.2020.2997255 (2169-3536)	100
2	Decision Boundary-Based Anomaly Detection Model Using Improved AnoGAN From ECG Data	IEEE ACCESS	정경용	8	미국	IEEE	SCI	08 June 2020	10.1109/ACCESS.2020.3000638 (2169-3536)	100
3	Optimization of Associative Knowledge Graph using TF-IDF based Ranking Score	Applied Sciences	정경용	10(13)	스위스	MDPI	SCI	2 July 2020	10.3390/app10134590 (2076-3417)	100
4	Word-Embedding-Based Traffic Document Classification Model for Detecting Emerging Risks Using Sentiment Similarity Weight	IEEE ACCESS	정경용	8	미국	IEEE	SCI	24 Sept 2020	10.1109/ACCESS.2020.3026585 (2169-3536)	100
5	Prediction of Traffic Congestion Based on LSTM Through Correction of Missing Temporal and Spatial Data	IEEE ACCESS	박찬홍	8	미국	IEEE	SCI	13 Aug 2020	10.1109/ACCESS.2020.3016469 (2169-3536)	100
6	Driver Facial Expression Analysis Using LFA-CRNN-Based Feature Extraction for Health-Risk Decisions	Applied Sciences	박찬홍	10(8)	스위스	MDPI	SCI	24 Apr 2020	10.3390/app10082956 (2076-3417)	100
7	Hybrid Multi-Modal Deep Learning using	IEEE ACCESS	정경용	8	미국	IEEE	SCI	16 Oct 2020	10.1109/ACCESS.2020.3031762	100

	Collaborative Concat Layer in Health Bigdata								(2169-3536)	
8	교통 영상 빅데이터 처리를 위한 Yolo 기반 광원 객체 탐지	융합정보 논문지	정경용	10(8)	대한민국	중소기업융합학회	비SCI	28 Aug 2020	10.22156/CS4SMB.2020.10.08.040 (2586-4440)	100
9	도로 위험 탐지를 위한 데이터 편향성 최적화 기반 연관 추론 모델	한국융합학회 논문지	정경용	11(9)	대한민국	한국융합학회	비SCI	28 Sept 2020	10.15207/JKCS.2020.11.9.001 (2233-4890)	100
10	Multi-Level Health Knowledge Mining Process in P2P Edge Network	IEEE ACCESS	정경용	9	미국	IEEE	SCI	16 Apr 2021	10.1109/ACCESS.2021.3073775 (2169-3536)	100
11	Associative Knowledge Graph Using Fuzzy Clustering and Min-Max Normalization in Video Contents	IEEE ACCESS	정경용	9	미국	IEEE	SCI	13 May 2021	10.1109/ACCESS.2021.3080180 (2169-3536)	100
12	Weight Feedback-Based Harmonic MDG-Ensemble Model for Prediction of Traffic Accident Severity	Applied Sciences	정경용	11(11)	스위스	MDPI	SCI	30 May 2021	10.3390/app11115072 (2076-3417)	100
13	Health Risk Detection and Classification Model Using Multi-Model-Based Image Channel Expansion and Visual Pattern Standardization	Applied Sciences	박찬홍	11(18)	스위스	MDPI	SCI	16 Sept 2021	10.3390/app11188621 (2076-3417)	100
14	Multimedia Knowledge-based Bridge Health Monitoring using Digital Twin	Multimedia Tools and Applications	강지수	80	미국	Springer	SCI	17 Feb 2021	10.1007/s11042-021-10649-x (1380-7501)	100
15	도로교통 이머징 리스크 탐지를 위한 AutoML과 CNN 기반 소프트 보팅 앙상블 분류 모델	융합학회 논문지	정경용	11(7)	대한민국	중소기업융합학회	비SCI	28 July 2021	10.22156/CS4SMB.2021.11.07.014 (2586-4440)	100
16	도로 CCTV 데이터를 활용한 딥러닝 기반 차량 이상 감지	한국융합학회 논문지	정경용	12(2)	대한민국	한국융합학회	비SCI	28 Feb 2021	10.15207/JKCS.2021.12.2.001 (2233-4890)	100

□ 국내 및 국제 학술회의 발표

번호	회의 명칭	발표 내용	발표자	발표 일시	장소	국명
1	The 10th International Conference on Convergence Technology in 2020(ICCT2020)	Risk Prediction using Correlation from Traffic Modality	백지원 구병국 류성은 박성수 정경용	8~10 July 2020	제주	대한민국
2	The 10th International	Traffic Knowledge Graph using Associative	김현진	8~10 July	제주	대한민국

	Conference on Convergence Technology in 2020(ICCT2020)	Document Weight	김민정 김주창 정경용	2020		
3	2020년도 한국인터넷정보학회 추계학술발표대회	교통안전을 위한 구조적 시각 유사도를 이용한 비지도 이상치 탐지	권혜정 김민정 강지수 백지원 정경용	30~31 Oct 2020	여수	대한민국
4	International Conference Convergence Content 2020(ICCC'2020)	Deep Autoencoder Model for Unsupervised Anomaly Detection in Traffic Safety	류성은 권혜정 김민정 정경용	12~14 Nov 2020	부산	대한민국
5	2021년도 한국인터넷정보학회 춘계학술발표대회	클러스터링 기반 중심적 분석을 이용한 실시간 교통안전등급 산출 모델	권혜정 김민정 류성은 정경용	23~24 Apr 2021	경주	대한민국
6	The Asia Pacific International Conference on Information Science and Technology 2021(APIC-IST'2021)	Extraction of Road Risk Factors Based on Regression Analysis to Prevent Traffic Accidents	김민정 권혜정 강지수 백지원 정경용	21 June 2021	부산	대한민국
7	The 11th International Conference on Convergence Technology 2021 (ICCT'2021)	Graph Neural Network based Recommendation Model for Traffic Safety	백지원 신동훈 정경용	1 July 2021	제주	대한민국
8	The 11th International Conference on Convergence Technology 2021(ICCT'2021)	Fast R-CNN based Imbalanced Posture Determination Model using OpenPose in Multi-Angle Image	권혜정 강지수 백지원 정경용	1 July 2021	제주	대한민국
9	2021년도 한국인공지능융합기술학회 하계학술대회	YOLO를 이용한 객체 지식 그래프 기반 이상 탐지	백지원 김민정 권혜정 강지수 정경용	26~27 Aug 2021	부산	대한민국
10	2021년도 한국인터넷정보학회 추계학술발표대회	BERT Bi-GRU 성능 향상을 위한 Gaussian 프로세스 기반 하이퍼파라미터 최적화	김민정 전병욱 권혜정 백지원 정경용	22~24 Oct 2021	정선	대한민국
11	2021년도 한국인공지능융합기술학회 추계학술대회	도로교통 이상 탐지를 위한 어텐션 마스킹 기반 적대적 학습	권혜정 전병욱 김민정 백지원 정경용	27 Nov 2021	인천	대한민국
12	The 13th International Conference on Internet 2021 (ICONI'2021)	Prediction of Anomaly Traffic Flow base on Distribution Distance using Semi-Supervised Learning	권혜정 전병욱 김민정 백지원 정경용	12~14 Dec 2021	제주	대한민국

□ 기술 요약 정보

번호	연도	기술명	요약 내용	기술 완성도	등록 번호	활용 여부	미활용사유	연구개발기관 외 활용 여부	허용방식

□ 보고서 원문

번호	연도	보고서 명	발간일	등록 번호

□ 생명자원(생물자원, 생명정보)/화합물

번호	생명자원(생물자원, 생명정보)/화합물 명	등록/기탁 번호	등록/기탁 기관	발생 연도

[기술적 성과]

□ 지식재산권(특허)

번호	구분	지식재산권 등 명칭 (건별 각각 기재)	국명	출원			등록			기여율	활용 여부
				출원인	출원일	출원번호	등록인	등록일	등록번호		
1	특허	비정형 데이터 기반 교통 위험 탐지 시스템 및 방법	대한민국	경기대학교 산학협력단	2021.04.12	10-2021-0046969				100	
2	특허	LSTM 기반 교통 혼잡도 예측 방법 및 장치	대한민국	경기대학교 산학협력단	2021.06.16	10-2021-0078113				100	
3	특허	특징 추출 기반의 LFA-CRNN을 이용한 운전자 표정 분석 장치 및 그 방법	대한민국	상지대학교 산학협력단	2021.08.19	10-2021-0109154				100	

ㅇ 지식재산권 활용 유형

번호	제품화	방어	전용실시	통상실시	무상실시	매매/양도	상호실시	담보대출	투자	활용 예정	기타
	ㅇ									ㅇ	

□ 저작권(소프트웨어)

번호	저작권명	창작일	저작자명	등록일	등록번호	저작권자명	기여율
1	빅데이터 기반 교통안전 이머징 리스크 마이닝 어플리케이션	2021.05.17	경기대학교 산학협력단	2021.06.17	C-2021-024308	경기대학교 산학협력단	100
2	이머징 리스크 예측을 위한 교통안전 지수 프로그램	2021.05.24	경기대학교 산학협력단	2021.06.17	C-2021-024309	경기대학교 산학협력단	100
3	지식기반 도심지 교통사고 위험도 예측 모바일 서비스	2021.09.16	상지대학교 산학협력단	2021.10.19	C-2021-041263	상지대학교 산학협력단	100

□ 신기술 지정

번호	명칭	고시일	보호기간	지정번호

□ 기술 및 제품 인증

번호	인증 분야	인증기관	인증 내용		인증 획득일	국가명
			인증명	인증번호		

□ 표준화

ㅇ 국내표준

번호	인증구분	인증여부	표준명	표준인증기구명	제안주체	표준종류	제안/인증일자

ㅇ 국제표준

번호	표준화단계 구분	표준명	표준기구명	표준분과명	의장단 활동여부	표준특허 추진여부	표준개발 방식	제안자	표준화 번호	제안일자

[경제적 성과]

□ 시제품 제작

번호	시제품명	출시/제작일	제작업체명	설치장소	이용분야	사업화 소요 기간	인증기관 (해당 시)	인증일 (해당 시)

□ 현장적용

번호	적용기술명	적용일	발주기관명	적용 연구개발기관명	적용현장명(장소)

□ 기술 실시(이전)

번호	기술 이전 유형	기술 실시 계약명	기술 실시 대상 기관	기술 실시 발생일	기술료	
					발생 연도	발생액(천원)
1	노하우	상황 딥러닝 기반 이머징 헬스 리스크 마이닝 노하우 기술	㈜코스티모어	2021.09.06.~2021.10.05	2021	2000

□ 사업화 투자실적

번호	연도	투자 자금 성격×	추가 연구개발 투자(천원)	설비 투자(천원)	기타 투자(천원)	합계(천원)

□ 사업화 현황

번호	사업화 방식	사업화 형태	지역	사업화명	내용	업체명	매출액		매출 발생 연도	기술 수명
							국내 (천원)	국외 (달러)		

□ 매출 실적(누적)

번호	사업화명	발생 연도	매출액		합계	산정 방법
			국내(천원)	국외(달러)		
합계						

□ 사업화 계획 및 무역 수지 개선 효과

성과					
사업화 계획	사업화 소요기간(년)				
	소요예산(천원)				
	예상 매출규모(천원)		현재까지	3년 후	5년 후
	시장 점유율	단위(%)	현재까지	3년 후	5년 후
		국내			
		국외			
	향후 관련기술, 제품을 응용한 타 모델, 제품 개발계획				
무역 수지 개선 효과(천원)	수입대체(내수)		현재	3년 후	5년 후
	수 출				

□ 고용 창출

번호	사업화명	사업화 업체명	고용창출 인원(명)		합계(명)
			yyyy년	yyyy년	
합계					

□ 고용 효과

구분			고용 효과(명)
고용 효과	개발 전	연구인력	
		생산인력	
	개발 후	연구인력	
		생산인력	

□ 비용 절감(누적)

번호	사업화명	발생 연도	산정방법	비용 절감액(천원)
합계				

□ 경제적 파급 효과

(단위 : 천원/년)

구분	사업화명	수입 대체	수출 증대	매출 증대	생산성 향상	고용 창출 (인력 양성 수, 명)	기타
해당 연도							
기대 목표							

□ 산업 지원(기술지도)

번호	내용	기간	참석 대상	장소	인원(명)

□ 기술 무역

(단위 : 천원)

번호	계약 연월	계약 기술명	계약 업체명	계약업체 국가	총 계약액	기 징수액	해당 연도	징수액	향후 예정액	수출/ 수입

[사회적 성과]

□ 법령 반영

번호	구분 (법률/시행령)	활용 구분 (제정/개정)	명 칭	해당 조항	시행일	관리 부처	제정/개정 내용

□ 정책활용 내용

번호	구분 (제안/채택)	정책명	관련 기관 (담당 부서)	활용 연도	채택 내용

□ 설계 기준/설명서(시방서)/지침/안내서에 반영

번호	구 분 (설계 기준/설명서/지침/안내서)	활용 구분 (신규/개선)	설계 기준/설명서/ 지침/안내서 명칭	반영일	반영 내용

□ 신진연구자 채용

번호	채용 연구개발기관명	채용일	근로기간	채용인원(명)	채용인원 중 청년인력(명)

□ 전문 연구 인력 양성

번호	분류	연도	현황(명)										
			학위별				성별		지역별				
			박사	석사	학사	기타	남	여	수도권	충청권	영남권	호남권	기타
1													

□ 산업 기술 인력 양성

번호	연도	프로그램명	프로그램 내용	교육 기관	교육 개최 횟수	총 교육 시간	총 교육 인원

□ 다른 국가연구개발사업에의 활용

번호	연도	중앙행정기관명	사업명	연구개발과제명	연구책임자	연구개발비(천원)

□ 국제화 협력성과

번호	구분 (유치/파견)	기간	국가	학위	전공	내용

□ 홍보 실적

번호	홍보 유형	매체명	제목	홍보일

□ 포상 및 수상 실적

번호	종류	포상명	포상 내용	포상 대상	포상일	포상 기관
1	Best Paper Award	우수논문상	The International Conference on Convergence Technology 2021에서 우수논문상 수상	주관-경기대학교 연구원	1 July 2021	한국융합학회
2	우수발표논문상	우수발표논문상	2021한국인공지능융합기술학회 추계학술대회에서 우수발표논문상 수상	주관-경기대학교 연구원	27 Nov 2021	한국인공지능융합기술학회

[인프라 성과]

□ 연구시설 · 장비

번호	구축 기관	연구시설/ 연구장비명	규격 (모델명)	개발여부 (○/×)	연구시설・장비 종합정보시스템× 등록여부	연구시설・장비 종합정보시스템× 등록번호	구축일자 (YY.MM.DD)	구축비용 (천원)	비고 (설치 장소)

2) 목표 달성 수준

연 구 개 발 목 표	달 성 내 용	달성도(%)
○(A-1)위험 예측을 위한 다중회귀기반 상황 심층 신경망 모델 개발	- 멀티-모달 Denoising 오토인코더 모델 - 결측치 추론을 통한 정확도 향상 - 협력적 레이어 기반 하이브리드 멀티-모달 딥러닝 - 연구결과를 **논문(SCI) 2편 게재** • IEEE Access 2020 8 104933-104943 게재 Multi-Modal Stacked Denoising Autoencoder for Handling Missing Data in Healthcare Big Data • IEEE Access 2020 8 192469-192480 게재 Hybrid Multi-modal Deep Learning using Collaborative Concat Layer in Health Bigdata - **성과 목표대비 논문(SCI) 1건 초과 달성**	100
○(A-2)상황 딥러닝 모델을 통한 이머징 리스크 상황 예측	- 상황 딥러닝을 통한 이머징 리스크 예측 - 의사결정 경계 기반 이상치 탐지 모델, 교통영상의 Yolo 기반 광원 객체탐지 - 연구결과를 **국내등재논문지(비SCI) 1편 및 SCI논문 1편 게재** • 융합정보논문지 2020 10(8) 40-46 게재 교통 영상 빅데이터 처리를 위한 Yolo 기반 광원 객체 탐지 • IEEE Access 2020 8(1) 108664-108674 게재 Decision Boundary-Based Anomaly Detection Model using Improved AnoGAN From ECG Data - **성과 목표대비 논문(SCI) 1건 초과 달성**	100
○(A-3)잠재적 위험에 따른 사용자별 교통안전지수 개발	- 순위스코어 연관지식 그래프 최적화 - 사용자별 잠재적 위험에 따른 교통안전지수 도출 - 연구결과를 **SCI논문 2편 게재** • Applied Sciences 2020 10(13) 4590-4610 게재 Optimization of Associative Knowledge Graph using TF-IDF based Ranking Score • IEEE Access 2020 8(1) 183983-183994 게재 Word Embedding based Traffic Document Classification Model for Detecting Emerging Risk using Sentiment Similarity Weight	100

	- **성과 목표대비 논문(SCI) 2건 초과 달성**	
○(C-1)교통혼잡도를 이용한 사고위험도 예측모형 개발	- LSTM 교통혼잡도 예측 및 정확도 향상 - 시계열 특성과 상황별 학습 및 예측 - 연구결과를 **SCI논문 1편 게재** • IEEE Access 2020 8(1) 150784-150796 게재 Prediction of Traffic Congestion Based on LSTM through Correction of Missing Temporal and Spatial Data	100
○(C-2)운전자 안면인식 기반 사고위험도 예측모형 개발	- 차원축소 및 운전자 안면인식 상태 판단, 도로교통 이머징 헬스 리스크 발견 - 데이터 편향성 최적화 연관 추론 모델 - 연구결과를 **국내등재논문지(비SCI) 1편 및 SCI논문 1편 게재** • 한국융합학회논문지 2020 11(9) 1-6 게재 도로 위험 탐지를 위한 데이터 편향성 최적화 기반 연관 추론 모델 • Applied Sciences 2020 10(8) 2956-2975 게재 Driver Facial Expression Analysis Using LFA-CRNN-Based Feature Extraction for Health-Risk Decisions - **성과 목표대비 논문(SCI) 1건 초과 달성**	100
○(B-1)사용자별 라이프로그 기반 헬스정보 분석 모델 개발	- 멀티스레딩을 통한 빅데이터의 병렬 수집 모듈 개발 - 멀티 레벨 헬스 지식 마이닝 프로세스 개발 - 연구결과를 **SCI논문 1편 게재** • IEEE Access 2021 9(1) 61623-61634 게재 Multi-Level Health Knowledge Mining Process in P2P Edge Network	100
○(B-2)시공간 결측 보정을 통한 LSTM 기반 교통위험 예측 모델 개발	- LSTM 딥러닝 모델 정확도 성능 고도화 - LSTM 기반 리스크 예측 모델, 교통사고 심각도 예측 고도화 - 교통사고 심각도에 영향을 주는 변수 예측 - 연구결과를 **특허 1건 출원 및 SCI논문 1편 게재** • 특허출원, 출원번호 10-2021-0078113, 2021.06.16. LSTM 기반 교통 혼잡도 예측 방법 및 장치 • Applied Sciences 2021 11(11) 5072-5089 게재 Weight Feedback based Harmonic MDG-ensemble Model for Prediction of Traffic Accident Severity - **성과 목표대비 논문(SCI) 1건 초과 달성**	100
○(B-3)빅데이터 기반 교통안전 이머징 리스크 마이닝 예측	- 데이터 편향성을 고려한 연관 마이닝 기반 도로위험 탐지 모델 - 숨겨진 연관정보에서 마이닝 기반 지식 그래프 발견 - 이상 상황에 대한 안전도 예측 모듈 개발 - MOG2를 이용한 배경 추출 및 딥러닝 기반 차량 이상 감지 알고리즘 개발 - 연구결과를 **SCI논문 2편 및 국내등재 학술지(비SCI) 1편 게재, SW등록 2건, 특허(출원) 1건** • IEEE Access 2021 9(1) 74802-74816 게재 Associative Knowledge Graph using Fuzzy Clustering and Min-Max Normalization in Video Contents • Multim. Tools Appl. 2021 80(26) 34609-34624 게재	100

	Multimedia Knowledge-based Bridge Health Monitoring using Digital Twin • 융합정보논문지 2021 11(7) 14-20 게재 도로교통 이머징 리스크 탐지를 위한 AutoML과 CNN 기반 소프트 보팅 앙상블 분류 모델 • SW저작권등록, 등록번호 C-2021-024308, 2021.06.17. 빅데이터 기반 교통안전 이머징 리스크 마이닝 어플리케이션 • SW저작권등록, 등록번호 C-2021-024309, 2021.06.17. 이머징 리스크 예측을 위한 교통안전 지수 프로그램 • 특허출원, 출원번호 10-2021-0046969, 2021.04.12. 워드임베딩을 이용한 감성 유사도 가중치 기반 교통위험 탐지 및 예측 모델 - **성과 목표대비 논문(SCI) 2건 초과 달성** - **성과 목표대비 논문(비SCI) 1건 초과 달성** - **성과 목표대비 특허출원 1건 초과 달성** - **성과 목표대비 SW등록 1건 초과 달성**	
○(D-1)딥러닝 기반 미래 교통 혼잡도 예측 시스템의 고도화	- 클라우드 기반 소셜 네트워크 구성 매커니즘 개발 - 프레임 전송간의 유휴시간(idle time) 감소 - 딥러닝 기반 교통 혼잡도 예측 시스템 성능 향상 - 연구결과를 **SCI논문 1편 및 국내등재 학술지(비SCI) 1편 게재, 특허(출원) 1건** • 한국융합학회논문지 2021 12(2) 1-6 게재 도로 CCTV 데이터를 활용한 딥러닝 기반 차량 이상 감지 • Applied Sciences 2021 11(18) 8621-8640 게재 Health Risk Detection and Classification Model using Multi-Model-based Image Channel Expansion and Visual Pattern Standardization • 특허출원, 출원번호 10-2021-0109154, 2021.08.19. 특징 추출 기반의 LFA-CRNN을 이용한 운전자 표정 분석 장치 및 그 방법 - **성과 목표대비 논문(SCI) 1건 초과 달성**	100
○(D-2)사고 위험도 예측 모바일 서비스 개발	- 가장 안전한 경로 추천 및 정보 시각화 모듈 개발 - 정체 예측정보 및 시간별 도로 위험도 점수 제공 모듈 개발 - 시각화된 정보를 통한 사용자의 이해 시간 감소 - 연구결과를 **SW등록 1건** - SW저작권등록 등록번호 C-2021-041263. 2021.10.19. 지식기반 도심지 교통사고 위험도 예측 모바일 서비스	100

4. 목표 미달 시 원인분석

○ 해당 없음

5. 연구개발성과의 관련 분야에 대한 기여 정도

○ 경제적·산업적 측면

- 본 제안에서 수행하는 연구를 통해 지속적으로 증가하는 도심 교통혼잡을 해결하기 위한 기술적 방안을 제시하여 교통혼잡비용에 따른 경제사회적 피해를 해결함
- 인공지능 및 빅데이터 기반의 예측정보 생성으로 교통정보의 정확도 향상 기술의 습득 및 관련 산업의 저

변 확대와 교통정보 관련 부가서비스의 질적 향상을 기대할 수 있음

- 교통정보 및 자동차산업과 연계한 새로운 부가가치 산업의 창출이 가능함
- 상황모션 추적 기반의 행동변화 모니터링과 상태감지를 이용한 3차원 위험 상황인식 시뮬레이션은 교통안전 실용화 제품개발이 가능함
- 빅데이터 마이닝 분야에서 우리나라의 국제적 위상의 고취, 마이닝의 기반 교통안전에 대한 지적재산권 확보를 할 수 있을 것으로 예상

○ 사회적 측면

- 교통물류에서 교통기술과 IT 융복합 기술을 통한 차세대 교통안전 발전에 기여
- 마이닝 기반 교통기술로 구체적인 활용이 가능하여 스마트 자동차, 자동차 안전기술, 자동차 편의기술, 융합기반 기술의 성장 및 발전에 기여
- 교통안전과 IT 융복합 기술을 통하여 국민의 실생활과 밀접하게 관련이 있는 교통사고 감소 기술개발로 사회적 비용 절감 및 국민 삶의 질 향상 기여
- 지속가능한 교통정책의 일환으로 미래형 도시부 도로에 대한 운영기반 마련
- 기초 학문과 응용학문이 서로 융합된 새로운 학제 간 복합연구를 위한 연구 및 교육기반 조성이 가능하며, 향후 다학제적 배경 하에 상황 딥러닝 기반 교통안전과 이머징 헬스 리스크 탐지 및 예측에 관한 종합적인 안목과 전문 인력을 양성할 수 있음
- 창의·도전적 연구결과를 고등교육과정에 접목하여, 전문화, 고도화, 세분화되어가는 실무영역에서 사용자를 고려하는 창의·융합에 대한 체계적, 실무적인 정보를 제공함으로써 미래지향적 사회요구에 대응이 가능함

○ 기술적 측면

- 교통안전, 인지과학, 창조과학, 정보과학, 상황인식 등의 창의도전 융복합 학문의 접근을 통한 새로운 학문적/기술적 패러다임으로 활용이 가능함
- 사용자별 위험도 노출에 따른 교통사고 유발 가능성이 있는 잠재적 위험도와 차량운전안전지수 서비스를 기상청에 제공하는 생활기상지수, 보건기상지수, 산업기상지수와 연계가 가능함
- 독창적인 접근 방식으로 스마트 헬스 기반 교통안전과 3차원 위험 상황인식은 국제표준 제정을 위한 데이터로 활용이 가능하고, 파급 효과가 있을 것으로 기대
- 사용자별 위험 상황인식을 활용하여 위험도 노출에 따른 **교통사고 유발 가능성이 있는 잠재적 위험도와 교통안전지수** 제공은 모바일 디바이스에서 가능함
- 사용자의 건강상태와 차량 주행 중의 외부 환경의 변화에 따른 상황 정보를 이용하여 컨텍스트 추론 기술 개발, 추론 결과를 이용한 개인화 서비스가 가능함

6. 연구개발성과 관리 및 활용 계획

○ 이머징 헬스 리스크 마이닝 기반 연구성과의 실용성 검증 방안

- 재난 재해 예방 등 안전관리 분야에서 **상황 딥러닝 기반 이머징 헬스 리스크 마이닝 모델**과 도심지 교통사고 위험도 예측 모델을 구체적으로 활용하는 융복합 연구의 지속적 추진이 필요함
- 스마트시티, 지능형 교통기술에서 기초 학문과 응용학문이 서로 융합된 새로운 학제 간 복합연구 및 교육기반 조성이 가능하며, 지역 맞춤형 교통사고 상황 및 예방에 관한 종합적인 안목과 전문 인력을 양성할 수 있음
- 창의·도전적 연구결과를 안전과 재난관리에 접목하여, 전문화, 세분화되어가는 도로교통 기술 분야에서 사용자를 고려하는 창의·융합에 대한 체계적, 실무적인 정보를 제공함으로써 미래지향적 사회요구에 대응이

가능함

○ 창의·도전적 연구의 결과물은 **교통사고 유발 네트워크 구축**을 위한 **토픽 모델링과 소셜 빅데이터 분석 기반 이머징 리스크 탐지 및 예측 기술을 개발**하여 교통사고 유발 가능성이 있는 잠재적 위험도를 판단. 위험 상황인식을 위한 지식베이스 분석 알고리즘 개발하고 잠재적 위험에 따른 사용자별 교통안전지수를 제공

- 교통사고 저감을 위한 라이프로그 기반 마이닝과 지식 발견을 이용한 연관 패턴 분석, 상황모션 추적 기반의 행동변화 모니터링, 상태감지를 이용한 3차원 위험 상황인식 시뮬레이션을 개발
- **사용자별 교통사고 잠재적 위험도와 교통안전지수를 멀티모달 인터랙션**으로 개발하여 스마트 헬스 기반의 교통안전을 제공하고 교통사고를 예방
- 교통안전산업에서 교통기술과 융복합 기술을 통해 잠재적 위험도에 따른 교통사고를 획기적으로 저감하고, 교통사고 사회문제 해결
- 지능형 도로교통체계와 위험 상황인식의 융합을 통한 시너지하여 상황 딥러닝 기반의 교통안전지수를 제공하여 국민 삶의 질 향상
- 차량 내의 환경인식센서, 영상센서, 모션센서 등을 설치하고, 위험 상황인식과 상황모션 추적 기반의 차량 운전자 모니터링을 하여 교통사고를 저감
- 교통산업과 IT융합 포럼의 활용 산/학/연 공동 활동을 통해 구체적인 협력 융복합 분야를 도출하고 중점 추진 분야와 분야별 구체화 가능

○ 사용자별 교통사고 유발 가능성이 있는 건강과 교통사고의 잠재적 위험도와 고도화된 교통안전지수를 제공하므로 교통사고에 대한 예방과 대응 관련 사회문제를 해결

- 국민의 실생활과 밀접하게 관련이 있는 **교통사고 예방 및 저감 기술로 사회적 비용 절감**하고 국민 삶의 질 향상에 기여
- **멀티모달 인터랙션 기반의 사용자 인터페이스**에서 잠재적 위험도와 교통안전지수를 단계별 주의사항과 대처 방안을 실시간으로 서비스 제공
- 3차원 위험 상황인식을 개발하여 도로교통 이용자별 차별화된 고도화 서비스 가능

주 의

1. 이 보고서는 국토교통부에서 시행한 국토교통과학기술 연구개발사업 '교통안전을 위한 상황 딥러닝 기반 이머징 헬스 리스크 마이닝' 연구개발과제 최종보고서이다.

2. 이 연구개발내용을 대외적으로 발표할 때에는 반드시 국토교통부(국토교통과학기술진흥원)에서 시행한 국토교통과학기술 연구개발사업의 결과임을 밝혀야 한다.

3. 국가과학기술 기밀 유지에 필요한 내용은 대외적으로 발표 또는 공개하여서는 안 된다.