creative commons

COMMONS DEED

저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:

저작자표시. 귀하는 원저작자를 표시하여야 합니다.

비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.

변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

**저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.**

이것은 이용허락규약(Legal Code)을 이해하기 쉽게 요약한 것입니다.

Disclaimer

dCollection

보건학 석사 학위논문

# 건강 지표 연관성을 활용한 웹 인터페이스 구축 및 질병 예측 연구

# PREDICTION OF DISEASE RISK AND WEB INTERFACE USING CORRELATED HEALTH INDEXES

**2018년 8월**

서울대학교 보건대학원

보건학과 생명정보학 전공

김 윤 정

# 국 문 초 록

## 건강 지표 연관성을 활용한

## 웹 인터페이스 구축 및 질병 예측 연구

서울대학교 보건대학원 보건학과
생명정보학 전공

**김 윤 정**

과학 기술의 발달과 생활수준의 향상 등으로 인간의 기대 수명은 전 세계적으로 꾸준히 증가하고 있다. 하지만 삶의 질을 증가시킬 수 있는 건강 수명을 연장하기 위해서는 식습관 변화 및 신체활동 감소 등으로 증가되고 있는 각종 질환의 발병률이나 유병률을 낮추는 것이 중요하다. 만성질환의 경우 지속적인 관리는 가능하지만 대부분 기능적 장애나 불구를 동반하게 되고, 대체로 연령의 증가와 더불어 증가하는 경향이 있으며, 예방 접종이나 약물 치료로 인한 완치가 어렵기 때문에 현재 전 세계적으로 가장 흔한 사망 및 장애 원인으로 알려져 있다. 또한 특정한 치료 방법이 없어 단일 질병 상태로 유지되기 보다는 시간이 지나면서 합병증을 유발하거나 건강 상태를 악화시켜 심각한 문제를 야기하게 되며, 이러한 합병증을 갖는 환자의 경우 단일 질병을 가진 환자보다 조기에 사망 할 확률이 높고 치료 및 병원 입원 기간이 길기 때문에 특히 면역에 취약한 노인층이나 저소득층에게 많이 나타나게 된다. 해외에서는 이미 지난 몇 년 전부터 많은 의료 기획자와 정부에서 이와 같은 문제성을 인지하고 연령에 관련된 질환 연구를 시행하고 있다. 국내에서도 개개인의 건강에 대한 관심도가 크게 증가하면서 관련 연구가 진행되

고 있지만, 시간이나 경제적인 제약들로 인해 의료 서비스를 받지 못하거나 적절한 치료시기를 놓쳐 질병을 악화시키는 경우가 증가하고 있다. 따라서 질병의 효과적인 관리를 위해서는 기존의 치료 질병 중심의 접근 방식에서 확장하여 개인의 행태에 따른 건강 상태를 스스로 점검하고, 발병 가능한 질환에 대한 사전 지식을 습득하여 예방 가능한 시스템이 이루어져야 한다. 본 연구에서는 통계청에서 제공하는 자료를 참고하여 국내 주요 사망 원인이 되는 질병들을 분류하고, 국민건강영양조사에서 제공하는 설문 조사 자료를 통해 주요 질병들의 지표로 활용될 수 있는 항목들을 선별하여 질환 관련 통합 데이터베이스인 LCBB-SC를 구축하였다. 구축한 데이터베이스를 활용하여 각종 질환들과 위험 요소들 사이의 연관성을 파악하기 위해 통계 및 머신 러닝 분석을 시행하였으며, 기존에 잘 알려진 질병 유발 요인들 외에도 잠재적 가능성이 있는 목록들을 관련도 수치와 함께 산출하여 새로운 요인 분석 연구의 기초 자료가 될 수 있는 정보를 제공하고자 하였다. 뿐만 아니라, 자가 증상 점검이 가능한 웹 인터페이스를 개방함으로써 본인의 건강 행태나 증상 정보 입력을 통해 발병 가능한 질환 목록을 제시해줄 수 있는 시스템을 구축하여 관심 질병에 대한 효율적인 정보 검색 및 접근이 가능하도록 하였으며 사전에 예방 가능한 질병 관리에 유용한 정보를 함께 제시하여줌으로써 개인의 건강 관리에 도움을 주고자 하였다. 이를 통해 궁극적으로 국내 주요 질병의 발병과 유병률을 낮추어 공중보건 향상에도 기여할 수 있을 것으로 기대된다.

주요어: 데이터베이스, 만성질환, 머신 러닝, 인공 신경망, 자가 진단 프로그램, 바이오인포매틱스

학 번: 2016-24002

*I dedicate this paper to my family & professor*

# 목 차

국문 초록 ····· i

목 차 ····· iv

표 목 차 ····· vii

그림 목차 ····· viii

약 어 ····· ix

**제 1 장. 서 론**

1.1 연구배경 ····· 1

1.1.1 고혈압 ····· 3

1.1.2 뇌졸중 ····· 3

1.1.3 당뇨 ····· 4

1.1.4 만성 폐쇄성 폐질환 ····· 5

1.1.5 악성 신생물 ····· 6

1.1.6 비만 ····· 7

1.2 인경 신공망 알고리즘 ····· 11

1.3 자가진단 시스템의 필요성 ····· 15

1.4 연구의 필요성 ····· 18

1.5 연구의 목적 ····· 22

## 제 2 장. 연구 방법

2.1 데이터 수집 및 가공 ····· 24
2.2 시스템 개발 환경 및 서버 구축 ····· 31
2.3 데이터베이스 구현 ····· 32
2.4 머신 러닝을 활용한 분석 ····· 35
2.4.1 머신 러닝과 라이브러리 ····· 35
2.4.2 신경망 구현 및 적용 ····· 37
2.4.3 클러스터링을 활용한 분석 ····· 40
2.5 알고리즘 성능 평가 ····· 43

## 제 3 장. 연구 결과

3.1 건강 지표 기반 질병 검색 시스템 ····· 46
3.2 머신 러닝을 활용한 분석 결과 ····· 51
3.2.1 머신 러닝을 활용한 연관성 분석 ····· 51
3.2.2 인공신경망을 활용한 분석 ····· 51
3.3 상관 분석과 거리 행렬 ····· 56

## 제 4 장. 고 찰

4.1 국내 연구 현황 ····· 57
4.2 질병 및 건강 지표의 연관성 ····· 59
4.3 연구의 한계점 ····· 66
4.4 활용 방안 및 기대 성과 ····· 68

**제 5 장. 결론 및 총론**

5.1 결론 ····· 70
5.2 총론 ····· 73

참고 문헌 ····· 76

Abstract ····· 86

# 표 목 차

**Table 1.1** The cause of death in South Korea (2015) ····· 9
**Table 1.2** Rank for top 20 causes of YLLs 1990~2015, South Korea ····· 10
**Table 1.3** Description of the studies concerning multimorbidity ····· 21
**Table 2.1** Selected risk factors and related symptoms for major chronic diseases ····· 26
**Table 2.2** Selected health behaviors that can have an indirect effect ····· 27
**Table 2.3** Selected major chronic diseases ····· 28
**Table 2.4** Other related factors and weight ····· 29
**Table 2.5** Type and number of data used for system construction ····· 30
**Table 2.6** Machin learning library for Java ····· 41
**Table 3.1** Multiple layer perceptron result for chronic disease (10-folds) ····· 54

# 그 림 목 차

**Figure 2.1** Database schema ····· 34
**Figure 2.2** Machine learning process ····· 42
**Figure 2.3** Multilayer Perceptron (MLP) algorithm ····· 45
**Figure 3.1** Main page of the constructed database ····· 48
**Figure 3.2 (a)** Example of self-symptom checker tool ····· 48
**Figure 3.2 (b)** Example of self-symptom checker tool ····· 49
**Figure 3.2 (c)** Example of self-symptom checker tool ····· 49
**Figure 3.3** Web interface result page ····· 50
**Figure 3.4** Provide major chronic disease information ····· 50

# 제 1 장. 서 론

## 1.1 연구배경

과학 기술의 발달과 생활수준의 향상 등으로 인간의 기대수명(life expectancy)은 전 세계적으로 꾸준히 증가하고 있다(World health statistics, 2015). 최근 산업 국가 35개국을 대상으로 한 통계 연구 결과에서 평균 기대 수명이 연장될 확률은 여성 65%, 남성 85%로, 모든 대상 국가에서 증가할 것으로 예상되었다(Kontis et al., 2017). 국내 통계청의 보고에 따르면 한국인의 기대수명 또한 2016년 82.4세로 지난 1970년에 비하여 약 20세가량 늘어난 것으로 나타났으며, 이러한 평균 수명의 연장은 노인인구의 증가 와도 연결되어 2026년 국내 노인 인구가 차지하는 비율이 20%가 넘게 되면서 초고령 사회에 진입하게 될 전망이다. 하지만 노인 인구의 증가와 더불어 식습관의 변화, 신체 활동의 감소 등으로 인한 만성질환 유병률(prevalence)의 증가가 사회적으로 중요한 문제로 대두 되고 있다.

만성질환은 오랜 기간 지속되는 질병 상태로, 주로 3개월 이상의 질병 경과가 나타나는 경우를 가리키며(Nodhturft et al., 2000) 장기간의 감독, 관찰 또는 돌봄이 필요하고 사람의 신체적, 심리적 및 사회적 기능을 방해하는 영구적이고 비가역적인 상태로 정의된다(Stuifbergen and Rogers, 1997). 급성 질환과 비교했을 때 대체로 병의 진행이 느리고 유병 기간이 긴 질환으로 증세가 완만하게 나타났다가 호전과 악화를 반복하며 점차 나빠지는 경향으로 간다. 지속적인 관리는 가능하지만 대부분 기능적 장애나

불구를 동반하게 되고, 대체로 연령의 증가와 더불어 증가하는 경향이 있으며 예방 접종이나 약물 치료로 인한 완치가 어렵기 때문에 현재 전 세계적으로 가장 흔한 사망 및 장애 원인으로 알려져 있다(Adams et al., 2013). 2015년 통계청 조사 자료에 의하면 한국인의 사망 원인 10위 중 7개가 만성질환으로 밝혀졌으며, 주요 만성질환으로는 고혈압(hypertension), 뇌졸중(stroke), 당뇨, 만성 폐쇄성 폐질환(chronic obstructive pulmonary disease, COPD) 악성 신생물, 비만이 있다<Table 1.1>. 뿐만 아니라 1995년과 2015년도의 조기 사망으로 인한 질병 부담(years of life lost, YLL)을 비교한 결과 대부분의 원인이 만성질환임을 확인할 수 있다< Table 1.2>.

만성질환의 직접적인 발생 요인은 아직 명확하게 밝혀진 것이 없으며, 여러가지 요인이 복합적으로 작용하여 발생하게 된다. 악성 신생물(malignant neoplasm)이나 당뇨(diabetes mellitus)의 경우 유전적 요인이 중요하게 작용하며 이 외에도 직업, 교육 수준, 거주지, 월 수입과 같은 사회 경제적 요인과 식습관, 규칙적인 운동 및 흡연, 음주, 약물과 같은 생활습관 요인과도 깊은 관련이 있는 것으로 알려져 있다(Lindstrom et al., 2006). 따라서 치료제에 의존하기 보다는 조기에 위험성을 인지하고 주기적인 건강 검진과 평소 식습관이나 생활 습관을 개선하여 질병을 사전에 예방하는 것이 중요하다. 1차 예방(primary prevention)은 질병에 걸리기 전 병의 원인이 되는 요소들을 사전에 제거하는 것으로, 질병이나 손상으로부터 건강한 사람을 보호하기 위한 단계이다(Gordon, 1983). 환경 위생, 식품 위생과 같은 공공 위생의 확보와 운동, 영양 조절, 금연, 절주 및 식습관 개선 등 개인의 건강증진에 목적을 두고 있다(Lindstrom et al., 2006). 2차 예방(secondary prevention)의 경우, 이미 질병에 걸린 상태에서 이를 조기에 발견하고 신속한 치료를 통해 질병의 악화를 방지하거나 완치에 목적을 두고있으며, 이를 위해서는 정기적인

신체 검사나 건강 검진이 중요하다(Gordon, 1983; Wagena et al., 2005). 3차 예방(tertiary prevention)은 관리의 마지막 단계로 신체의 퇴보를 방지하고 삶의 질을 극대화하여 후유증 및 합병증을 최소화 하고 재활을 통한 사회 복귀 및 적응에 목적을 두고 있다(Gordon, 1983).

### 1.1.1 고혈압

고혈압의 경우, 가장 흔하지만 관리가 어려운 퇴행성 질환으로 뇌졸중, 심근경색증 등 심뇌혈관질환의 발생과 관련이 높다고 알려져 있다(Cosentino and Volpe, 2005). 고혈압의 종류로는 원인이 불명확하며 고혈압 환자의 약 95%를 차지하고 있는 본태성 고혈압(essential hypertension)과 심혈관 질환, 신장 질환, 신경 질환, 부신 질환, 갑상선 질환 등에 의해 발생하는 2차성 고혈압(secondary hypertension)이 있다(Carretero and Oparil, 2000). 아직까지 명확하게 밝혀진 발생 기전은 없지만 주요 원인으로는 가족력, 음주, 흡연, 당뇨, 고지혈증, 노화, 운동 부족, 비만, 염분과 지방질의 과다섭취, 스트레스 등과 같은 심리적 및 환경적 요인을 들 수 있다(Chobanian et al., 2003). 본태성 고혈압의 경우는 혈압을 140/90mmHg 이하로 유지하며 노인의 경우는 150mmHg이하로 조절하는 것이 일차적 목표이다. 경증 고혈압의 경우 규칙적인 운동, 체중 감소, 염분 제한 및 칼슘, 포타슘 섭취 증가, 금주와 같은 방법으로 약물 요법 없이 3개월 정도 시행할 수 있으나 중증도 이상의 경우에는 약물요법과 병행 해야 한다(Carretero and Oparil, 2000).

### 1.1.2 뇌졸중

뇌졸중은 뇌혈관이 막히거나 파열되어 뇌기능의 일부 혹은 전체의 기능을 상실하여 그 장애가 상당 기간 지속되는 질병으로

크게 허혈성 뇌경색과 뇌출혈로 나눌 수 있다(Sacco et al., 2013). 허혈성 뇌경색의 경우 뇌 혈관이 혈전이나 색전에 의해 막혀서 발생하며, 뇌혈관이 파열되어 뇌 조직 내부로 혈액이 유출되어 뇌 조직을 압박하는 경우 뇌출혈(cerebral hemorrhage)이라고 한다(Hylek and Singer, 1994). 뇌졸중의 주요 원인으로는 동맥경화, 고혈압, 고지혈증 등의 혈관 질환이 있으며, 두통, 구토, 시력장애, 인지기능 저하, 사지 근력 저하, 편측 마비 안면 마비, 구음 장애와 같은 증상을 보인다(Mouradian et al., 2002). 평상시 고혈압, 당뇨, 고지혈증과 같은 만성질환 관리가 중요하며 이 외에도 음식조절, 주기적 운동, 금연, 절주 등의 건강증진 운동의 생활화가 예방에 도움이 된다(Sacco et al., 2013). 이미 발생한 경우에는 신속한 응급 처리로 사망과 후유증이 발생하지 않도록 조치가 필요하며, 후유증으로 인한 부담 해소 및 악화되지 않도록 재활 프로그램 등을 통한 관리가 이루어져야 한다(Adams et al., 1993).

### 1.1.3 당뇨

당뇨는 췌장의 인슐린 분비량이 부족하거나 정상적인 기능이 이루어지지 않아 혈중 포도당 농도가 높아져 소변으로 다량의 포도당을 배출시키는 대사 질환으로, 혈당이 공복 시 126mg/dl 혹은 식후 2시간 후 200mg/dl이상 되는 경우 당뇨로 정의 된다(American Diabetes Association, ADA, 2010). 제 1형 당뇨의 경우 자가 면역 반응으로 인해 췌장 랑게르한스 섬의 베타세포가 파괴되어 인슐린이 분비되지 않아 따로 인슐린 투여가 필요하며, 소아나 청소년기에 주로 발생하여 소아 당뇨병 이라고도 한다. 제 2형 당뇨의 경우 췌장에서 인슐린이 정상적으로 분비 되더라도 신체에 대한 반응성이 떨어져서 인슐린이 기능을 발휘하지 못하는 경우로 전체 당뇨 환자의 80% 이상이 여기에 속한다(Alberti and Zimmet,

1998). 현재까지 밝혀진 원인으로는 유전, 과체중, 비만, 운동부족, 스트레스 등의 요인과 췌장 수술, 감염, 약제 임신 등이 있으며 주로 뇨, 다식, 체중 감소와 같은 증상이 나타난다(Stamler et al., 1993; Lindstrom et al., 2006). 당뇨의 경우 유전적인 요인을 제외하고는 사전에 관리를 통해 발생 시기를 늦추거나 당뇨병으로 인한 합병증을 예방하는 것이 가능하다. 당뇨 판정을 받더라도 운동, 체중조절, 식이요법 등으로 인슐린을 분비하는 췌장 베타 세포의 기능을 보존하고 인슐린의 저항성을 낮추어 정상 혈당 범위를 유지할 수 있다(Ramachandran et al., 2006).

### 1.1.4 만성 폐쇄성 폐질환

만성 폐쇄성 폐질환은 폐가 손상되어 시간이 지남에 따라 기도 저항이 증가하여 호기 시 늑강의 압력 증가로 인한 기도의 압박 때문에 점점 호흡이 힘들어지는 질환으로 폐 기능 검사 결과 측정된 기류가 불량할 때 진단 내려진다(Pauwels et al., 2001). 다른 만성질환들과 마찬가지로 초기에 뚜렷한 증상이 나타나지 않으며 주로 가래를 동반한 만성 기침이 가장 초기에 발생하여 호흡곤란으로 까지 이어지게 된다(Vestbo et al., 2013). 폐기종(emphysema)이나 만성기관지염(chronic bronchitis)의 경우 흡연이 가장 주 원인으로 밝혀져 있으며 이 외에도 분진, 대기오염, 유전, 잦은 알레르기 노출로 인해 기도나 폐에 만성 염증을 초래하여 기도 협착이나 폐 조직의 파괴로 호흡곤란을 겪게 된다(Silverman et al., 1998; Wagena et al., 2005) 급성 폐질환의 경우 감염이나 외부 오염 물질, 약물 오남용과 같은 요인에 의해 유발되며, 특히 건조하기 기온이 찬 겨울에 더 많이 발생할 수 있다. 또한 허혈성 심 질환, 고혈압, 당뇨, 폐암 과 같은 다른 만성질환을 동반하는 경우가 많다(Mannino et al., 2008). 대부분의 만성 폐쇄성

질환의 경우 흡연이나 간접흡연에 의한 연기 노출을 줄이고 대기의 질은 개선함으로써 예방이 가능할 수 있다. 미생물 감염에 의한 폐질환의 경우 예방접종을 통해 예방이 가능하다.

### 1.1.5 악성 신생물

악성 신생물 또는 암은 단일 질병이 아니라 정상적인 세포분열 조절 기작을 상실한 세포의 존재로 특성화 되는 이질적인 질병이다(Dorland, 2011). 정상 조절 기작에 대한 반응이 상실되면 암세포는 점차 규칙적인 형태와 경계 부위를 잃게 되고 결국 비정상 세포로 구성된 뚜렷한 세포 덩어리인 종양을 형성하게 된다. 종양을 구성하는 세포가 특정부위에 제한적으로 위치하고 있는 경우 양성종양(benign tumor)이라고 하며, 만일 세포들이 다른 조직을 침유하게 되면 악성종양(malignant tumor)라고 한다. 몸속의 다른 장소로 이동하는 세포들은 전이성 종양을 형성하게 되며, 이런 과정을 전이(metastasis)라고 한다(England et al., 1989). 악성 종양은 피부나 점막과 같은 상피 세포에서 발생하는 암종(carcinoma)과 근육, 결합조직, 뼈, 연골, 혈관 등 비 상피 세포에서 발생하는 육종(sarcoma)으로 나눌 수 있다(Greene et al., 2002). 환자가 가진 증상이나 의사의 객관적인 소견, 각종 혈액검사, 암 표지자 검사 등을 통한 임상 진단(clinic diagnosis)과, CT (Computed Tomography), MRI (Magnetic Resonance Imaging), 일반 X-ray, PET-CT (Positron Emission Tomography–Computed Tomography, 초음파, 뼈 스캔, 내시경 등의 이미지를 통해 진단하는 영상 진단(imaging diagnosis)이 있으며, 조직 검사나 수술을 통해 암세포를 직접 확인하는 조직 진단(pathologic diagnosis)을 통해 암을 확진할 수 있다(Pisano et al., 2005). 정확한 암의 발생 원인은 아직까지 확실하게 밝혀진 것이 없지만 근본적으로 유전적 요인과 방사선, 대기오염, 흡연 음주,

식이 등의 외적 요인에 영향을 많이 받는 것으로 알려져 있다. 한 연구 결과에 따르면 전 세계적으로 특정 암종의 발생 빈도가 지역별로 조사한 결과 다른 나라로 이주한 집단에서의 암 발생빈도는 정착한 나라의 암 발생빈도와 유사하게 나타난다는 결과를 보여주었으며, 이를 통해 암 발생 과정에서 환경적 요인의 중요성을 유추해볼 수 있다(Akazaki and Stemmermann, 1973). 암의 경우 다른 만성질환들과 비교 했을 때, 초기에 발견하는 경우 후기의 암에 비해 약물 반응도가 훨씬 높게 나타났으며 수술만으로 완치가 가능하고 재발률도 낮기 때문에 조기 진단의 중요성이 특히 강조 되고 있다(Casanovas et al., 2005).

### 1.1.6 비만

비만은 에너지 섭취와 소비의 불균형으로 인한 1차성 (primary obesity) 비만과 특정 질병이나 약물 등 다른 원인에 의해 2차적으로 발생하는 2차성 비만(secondary obesity)으로 나눌 수 있다(Lyznicki et al., 2001). 대부분의 비만 환자의 경우 1차성에 해당되며, 유전적 요인과 나이, 성별, 만성 스트레스, 고 열량의 식사 및 음주, 신체활동의 부족이 주요 원인으로 밝혀져 있다. 2차성 비만은 에너지 섭취량의 증가나 활동량의 감소가 없음에도 체중이 증가하는 경우, 체중감량이 쉽게 이루어지지 않을 때 의심해 볼 수 있다(Lyznicki et al., 2001). 항 정신병 약물, 항 우울 제, 당뇨 치료제, 스테로이드제, 경구 피임약, 각종 호르몬, 항 히스타민제에 의한 경우가 가장 흔한 원인으로 밝혀져 있으며, 이 외에도 갑상선 기능저하, 쿠싱 증후군(Cushing’s syndrome), 다낭성 난소 증후군 등과 같은 신경 및 내분비계의 질환에 의한 비만이 있다(Luppino et al., 2010; Yazdi et al., 2015). 대표적인 비만의 합병증으로는 인슐린 저항성과 관련이 있는 제 2형 당뇨, 이상지질혈증, 고혈압, 협심증, 심근경색증, 지방간,

담낭 질환, 수면 무 호흡, 통풍, 골 관절염, 대장암, 유방암, 난소 암 등이 있고 심리적인 질병과도 연관이 있다. 비만으로 인해 나타나는 동반 질환들은 체중 감량만으로도 예방 할 수 있기 때문에 정확한 체지방을 통해 비만도를 평가하고 사전에 예방하는 것이 중요하다(Kim, 2014).

**Table 1.1 The cause of death in South Korea (2015)**

| Rank | Cause of death | No. of death | Composition | Mortality(100K) |
|---|---|---|---|---|
| 1 | Neoplasms | 79,729 | 28.3 | 149.0 |
| 2 | Cerebrovascular diseases | 25,447 | 9.6 | 50.3 |
| 3 | Cardiovascular diseases | 25,365 | 9.5 | 50.2 |
| 4 | Self-harm | 14,427 | 5.4 | 28.5 |
| 5 | Diabetes | 10,888 | 4.1 | 21.5 |
| 6 | Pneumonia | 10,809 | 4.1 | 21.4 |
| 7 | Chronic respiratory | 7,074 | 2.7 | 14.0 |
| 8 | Liver cancer | 6,665 | 2.5 | 13.2 |
| 9 | Transport injuries | 6,028 | 2.3 | 11.9 |
| 10 | Hypertension | 4,732 | 1.8 | 9.4 |

The status of chronic diseases in Korea is still high despite the steady decline in standard mortality rate. Chronic diseases (highlighted) account for 81% of all deaths, and seven of the top ten causes of death are chronic diseases. Major chronic diseases include neoplasms, cerebrovascular diseases, cardiovascular diseases, diabetes, pneumonia, chronic respiratory, liver cancer and hypertension. In particular, mortality from diabetes and cerebrovascular disease was higher than the other OECD countries, suggesting that prevention is weak (Statistic Korea, http://kostat.go.kr/portal/korea/index.action).

**Table 1.2 Rank for top 20 causes of YLLs 1990~2015, South Korea**

| Rank of disorder, 1990 | No. of YLLs in thousands (total %) | Rank of disorder, 2015 | No. of YLLs in thousands (total %) |
|---|---|---|---|
| **Stroke** | **1,047 (15.9%)** | **Stroke** | **671 (12.6%)** |
| Road injuries | 572 (8.7%) | Self-harm | 425 (8.0%) |
| **Ischemic heart disease** | **549 (8.4%)** | **Ischemic heart disease** | **342 (6.4%)** |
| **Cirrhosis** | **414 (6.3%)** | **Liver cancer** | **338 (6.2%)** |
| **Stomach cancer** | **414 (6.3%)** | **Stomach cancer** | **329 (6.2%)** |
| **Liver cancer** | **285 (4.3%)** | **Lung cancer** | **313 (5.9%)** |
| Self-harm | 234 (3.9%) | **Cirrhosis** | **263 (4.9%)** |
| **Congenital anomalies** | **197 (3.0%)** | Road injuries | 249 (4.7%) |
| **Tuberculosis** | **178 (2.7%)** | **Diabetes** | **222 (4.2%)** |
| **Lung cancer** | **175 (2.7%)** | **Colorectal cancer** | **163 (3.1%)** |
| **Low respiratory infect** | **138 (2.1%)** | **Low respiratory infect** | **110 (2.1%)** |
| **Diabetes** | **138 (2.1%)** | **COPD** | **99 (1.9%)** |
| Drowning | 117 (1.8%) | **Pancreatic cancer** | **88 (1.7%)** |
| Mechanical forces | 117 (1.8%) | Falls | 83 (1.6%) |
| Falls | 100 (1.5%) | **Alzheimer's disease** | **70 (1.3%)** |
| **Leukemia** | **94 (1.4%)** | **Gallbladder cancer** | **68 (1.3%)** |
| **COPD** | **74 (1.1%)** | **Chronic kidney disease** | **64 (1.2%)** |
| **Asthma** | **69 (1.1%)** | **Tuberculosis** | **62 (1.2%)** |
| **Hypertensive hear disease** | **62 (1.0%)** | **Leukemia** | **59 (1.1%)** |
| **Colorectal cancer** | **61 (0.9%)** | **Hypertensive hear disease** | **57 (1.1%)** |

Comparing major causes of years of life lost in Korea between 1990 and 2015; most of the major causes of death are chronic diseases (bolded). The death rate due to external injuries was reduced, but the mortality rate due to chronic diseases did not decrease rather the proportion was attributed to the cause of premature death (WHO, http://www.who.int/en/).

## 1.2 인공 신경망 알고리즘

기존의 보건 의료 분야의 정보들이 전산화 과정을 통해 방대한 양의 데이터도 축적 되면서 초기 질병 탐지, 환자 치료 및 지역 사회 서비스와 같은 다양한 분야에서 이용되고 있다. 하지만 데이터의 품질이 불완전한 경우 분석 정확도가 떨어지게 되며 너무 많은 양의 데이터를 이용할 시 오히려 예측을 약화시킬 수 있다. 따라서 이를 효과적으로 다루기 위해서는 기존의 분석방법 보다는 기계 학습 알고리즘을 통해 간소화시켜주는 것이 중요하다. 본 연구에서 사용된 인공 신경망 알고리즘은 인간의 두뇌 신경 세포 구조를 모방하는 알고리즘으로, 원래는 뉴런의 전기 자극을 숫자로 나타내는 지도 학습의 일종으로 사용되었다(Smith and Bolam, 1990). 초기에 메커니즘을 재현하기 위해서 단순히 몇 개의 망 층을 쌓아 올리는 구조로 설계 되었지만 신경망이 인식할 수 있는 패턴에 제한이 많아지게 되면서 좀 더 복잡한 구조로 망의 링크를 수치화 시킨 가중치 값을 설정해주었다. 가장 처음 발표된 신경망 알고리즘은 퍼셉트론(perceptron)으로 1957년 Frank Rosenbelt에 의해 발표 되었다. 퍼셉트론 알고리즘은 신경망 알고리즘 중에서 구조가 가장 단순하며, 두 개의 범주에 대한 선형 분류를 할 수 있다(Rosenblatt, 1958; Memisevic, 2015). 퍼셉트론에 대한 일반적인 출력은 다음과 같이 나타낼 수 있다.

$$y(x) = f(w^T x)$$

$$f(a) = \begin{cases} +1, a \geq 0 \\ -1, a < 0 \end{cases}$$

f(*)는 계단 함수(step function)로 특징 벡터(feature vector)의 각 인자에 가중치를 곱해 그 합을 구하고, 계단 함수를 통해 활성화시켜

출력을 반환 한다(Kaluza, 2016). 학습이 진행되는 동안 이 결과를 실측 데이터와 비교하여 오차를 피드백 하면서 예측력을 높일 수 있다. 오차 함수를 최소화하기 위해 gradient descent 와 같은 최적화 알고리즘이 사용하여 local minimum 을 찾는다(Ng et al., 1997).

$$w^{(k+1)} = w^{(k)} - \eta \nabla E(w) = w^{(k)} + \eta \sum_{n \in M} x_n t_n$$

여기서 $\eta$ 은 학습률(learning rate)이며, 최적화 알고리즘에서 학습 속도를 조절하는 공통 파라미터고, $k$ 는 알고리즘 스텝을 나타낸다(Ng et al., 1997). 퍼셉트론의 경우 데이터 셋을 선형적으로 분리할 수 있으면 학습률 값에 상관없이 알고리즘이 항상 수렴한다는 것이 증명 되어있기 때문에 학습률 값을 1 로 설정한다. 일반적으로 머신 러닝 방법의 성능은 confuse matrix 를 기반으로 한 다음 세 가지 지표로 측정된다(Caballero et al., 2010).

$$\text{Accuracy} = \frac{number\ of\ true\ positives + number\ of\ true\ negatives}{true\ positives + flase\ positive + false\ negative + true\ negative}$$

$$\text{Precision} = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + number\ of\ false\ positives}$$

$$\text{Sensitivity} = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + number\ of\ false\ negatives}$$

정확도(accuracy)는 데이터 전체에 대비 옳게 분류된 데이터의 비율을 의미하고, 정밀도(precision)는 positive 로 예측된 데이터 대비 실제 positive 로 예측된 데이터의 비율을 의미한다. 민감도(sensitivity)는 실제 positive 인 데이터 대비 positive 로 예측된 데이터의 비율을 의미한다(Ramana et al., 2011; Caballero et al., 2010).

하지만 퍼셉트론은 선형으로 분류할 수 있는 문제에 대해서는 효율적이지만, 비선형 문제는 전혀 풀지 못한다는 단점이

있다(Parlos et al., 1994). 대부분의 분석에 사용되는 문제는 비선형이기 때문에 이에 대응하기 위해 다층 퍼셉트론(MLP) 혹은 다층 신경망(multi-layer neural networks)으로 개선되었다. 입력층(input layer)과 출력층(output layer)사이에 은닉층(hidden layers)을 추가함으로써, 다양한 패턴 분석이 가능하다. 이때 은닉층을 통하지 않고 바로 입력층에서 출력층으로 직접 가는 연결(skip-layer connection)이 생기지 않아야 한다(Chen et al., 1990). 신경망에서는 이론상으로나 구현 상의 측면에서 신경망들의 유닛들 사이의 링크에 사이클이 없는 구조의 모델(feed-forward structure)를 갖게 하는 것이 좋으며 이러한 패턴으로 은닉층 수를 늘려가며 임의의 함수를 근사(approximation) 할 수 있다(Gardner and Dorling, 1998). 다층 퍼셉트론을 계산하는 방법은 다음과 같다.

$$E(W,b) = -lnL(W,b) = -\sum_{n=1}^{N} E_n(W,b) \sum_{k=1}^{K} t_{nk}\, lnY_{nk}$$

이때 발생하는 오차를 입력층과 은닉층에 모두 반영하기 위해서 역전파(backpropagation) 알고리즘을 통해 가중치와 바이어스(bias)를 계산하여 신경망의 정확도를 높일 수 있다. 출력의 오차를 전체 망에 역전파함으로써 각 반복 마다 모델이 학습 데이터에 맞게 갱신되고 조정되어 데이터에 대한 최적화된 함수를 찾을 수 있다(Hinton and Salakhutdinov, 2006). 오차 함수의 값을 구하는 공식은 다음과 같다.

$$E(W,b) = \sum_{n=1}^{N} E_n(W,b)$$

일반적으로 학습률이 작을수록 알고리즘이 local minimum 에 빠질 가능성이 높아지는 반면, 학습률이 너무 커지면 값의 변동폭이 커지기 때문에 수렴하는 모델을 찾을 수 없게 되므로 처음에는 학습률을 크게 설정했다가 각 반복 마다 점차 줄여 나가야 한다. 망의 구조가 단순하다면 학습 데이터를 통해 가중치가 제대로 분배되어 패턴들을 잘 인식하고 분류하는 것이 가능하지만, 망이 복잡해지면 링크가 너무 조밀하게 많아지게 되고 가중치들 간의 차이를 만들기 어려워지는 단점이 생긴다(Dreiseitl and Ohno-Machado, 2002). 또한 학습과정에서 발생하는 오차들을 망 전체에 피드백시키는 메커니즘을 채택함으로써 올바른 모델을 만들어 가는 신경망은 망의 구조가 단순할 경우 피드백이 제대로 반영되지만 층이 많아지고 복잡해질수록 오차가 망 전체에 반영되기 전에 사라져 버리는 문제가 발생하게 된다(Karayiannis and Venetsanopoulos, 2013). 하지만 최근 딥러닝의 출현으로 여러 종류의 신경망 알고리즘이 소개되었으며, 그중 일부는 비지도 학습(unsupervised learning)으로 학습 과정에서 가중치들을 조정함으로써 예측률을 높일 수 있다.

## 1.3 자가진단 시스템의 필요성

고령화 시대에 접어들면서 개개인의 건강에 대한 관심도가 크게 증가하고 있으며, 전 세계적으로 이와 관련된 많은 연구가 진행 중이다. 특히 4차 산업 혁명의 주요 기술인 정보통신(ICT)의 급격한 발전으로 이를 활용한 온라인 자가 진단과 같은 간편한 시스템이 크게 각광받고 있다. 해외에서는 이미 많은 환자들이 신뢰할 만한 자가 진단 온라인 소스들을 이용하여 그들의 건강 상태와 관련 정보를 얻고 있다. 일부 서비스는 이러한 개인 증상 기록 정보를 전문의에게 전달하여 이를 의료 기록에 통합시켜 환자가 방문하기 전에 상태 파악이 가능하게 되어 시간을 절약하고 상담의 생산성을 높일 수 있다. 또한 의사들이 고려하지 못했던 질병에 대한 목록까지 보여주기 때문에 범위가 넓어져 진단에 도움을 줄 수 있다는 장점이 있다.

Family doctor (http://familydoctor.org/familydoctor/en/health-tools/search-by-symptom.html)는 미국 가정 의학 아카데미(American Academy of Family Physicians, AAFP)에서 개발한 자가진단 점검 툴로 사용자는 웹사이트에서 제공하는 몇 가지의 목록 중 본인의 증상을 선택 함으로써 쿼리를 시행할 수 있으며 이후 선택 값에 따라 추가 질문이 주어지거나 의심 가능 질병 목록을 제시하여 준다. Healthy Children symptom checker (http://www.healthychildren.org/english/tips-tools/symptom-checker)의 경우, American Academy of Pediatrics (AAP) 사이트에서 이용 가능하며, 화면에 표시된 아이 그림에 마우스를 갖다 대고 부위 별 증상을 선택하면 예상되는 질병 목록을 보여준다. 또한, 처음 아이를 키우는 부모에게 유용한 간단한 생활 건강 지식과 응급 상황에서도 대처 할 수 있는 가이드 라인을 함께 제시해 주고있다. Isabel Symptom Checker

(http://symptomchecker.isabelhealthcare. com/home/default)는 영국 charity Isabel Healthcare에서 기존의 서비스 이용자들에게만 제공되었던 증상 점검 시스템이었지만, 2012년 이후 대중에게 무료로 배포되면서 많은 이용자를 확보하고 있다. 약 6,000여 개의 질병을 다루는 이 시스템은 사용자의 나이, 성별, 임신 상태, 거주지와 같은 기본 정보들을 받고 일반적인 질병 정보와 함께 즉각적인 치료의 유무를 같이 판단하여 결과를 제시해 준다. 또한, 질병 별 전문 의학 분야 정보를 제공해주고 있어 이용자로 하여금 병원에 가기 전 효율적인 진료를 받는데 도움을 줄 수 있다. iTriage Symptom Checker (https://www.itriagehealth.com/) 는 2008년 2명의 응급 의학 의사가 설립한 의료 기술 회사인 iTriage에서 제공하는 서비스로, 사용자는 아바타에서 신체 일부를 선택하여 쿼리를 시행한다. 다양한 증상 목록을 사용하여 세밀한 점검이 가능하며, 이용자의 동의 하에 결과를 하버드 의과대학팀에 전달해줌으로써 리뷰 과정을 통해 추후 좀 더 정확한 결과를 도출해낼 수 있도록 업데이트가 가능한 시스템이다. Mayo Clinic Symptom Checker (http://www.mayoclinic.com/health/symptom-checker/DS00671)의 경우, 현재 미국 내에서 가장 많은 이용자를 확보하고 있는 Mayo Clinic에서 제공되는 서비스로 제한된 증상 목록 중 가장 관련성이 높은 증상을 선택하면 이와 연관된 세부 증상을 선택해 나가면서 점차 좁은 범위의 세분화된 질병 목록을 얻을 수 있다. 온라인 환자 교육 자료에서 사용된 유사 증상을 가진 환자들의 정보를 바탕으로 추가 정보를 얻을 수 있으며, 마찬가지로 관련 질병에 관한 전문적인 지식을 함께 제공해 이용자의 편의를 돕고 있다. WebMD (http://symptoms.webmd.com/) 또한 마찬가지로 사용자의 성별과 나이를 정보를 입력 받은 다음 증상이 나타나는 신체부위를 선택하고 진단 범위를 좁히기 위해 일련의 질문에 답을 하면 관련

질병 정보를 제공해준다. 또한 각 질병 별로 관련 논문이나 진단에 인용된 정보를 얻은 출처를 마지막 검토 시기와 함께 제공해 줌으로써 결과의 신뢰도를 높여 많은 이용자를 보유하고 있다.

이처럼 해외에서는 이미 자가 점검 시스템이 일반인들에게 개방되어 사용될 뿐 아니라 실제 의사들의 진단 시에도 참고 자료로 활용되고 있다. 현재까지 온라인 자가 점검 시스템의 효과에 대한 분석 결과는 뚜렷하게 나타나진 않았지만, 영국의 한 연구 결과에 의하면 이들은 자가 점검 시스템을 이용하여 실제 환자의 진단에 사용한 결과 70%의 정확도를 나타냈다고 밝혔다(Farmer et al., 2011). 또 다른 연구에 의하면, 뇌경색과 같이 복잡한 요인이 작용하는 질병의 경우 일반 통계자료를 활용하여 예측도 평가를 했을 경우 정확도가 50%에도 못 미쳤지만, 머신 러닝 알고리즘을 용합시켜 정확도가 94.80%에 도달하는 결과를 이루어 냈다. 국내에서도 2000년대 후반부터 많은 이와 관련된 많은 연구가 진행되어왔다. 한 국내 연구팀에 의해 ART2 기반의 지능형 자가 건강 진단 시스템이 제안되었다. 하지만, 자가진단 시스템에 적합하지 않은 유클라디안 거리 법을 적용하여 유사도를 측정하였기 때문에 진단 정확도가 떨어진다는 한계점이 있다. 또한, 이 시스템에서 사용한 일방적인 질문 방식은 사용자가 느끼지 못하는 증상들을 놓칠 수 있어 질병 도출의 정확성 역시 떨어진다는 문제점이 있다(Kim, 2014).

## 1.4 연구의 필요성

국민 건강 보험 공단과 건강 보험 심사 평가원에서 공동 발간한 건강보험통계연보에 따르면 국내 만성질환 진료자 수가 2006년 938만 명에서 2015년 1,439만 명으로 10년 사이 약 54% 증가하였으며, 현재까지도 매년 꾸준히 증가하는 것으로 나타났다. 또한, 만성질환에 의한 국내 사망 인구수가 전체 사망자 수의 81%를 차지하고 있으며, 주요 사망원인의 대부분이 만성질환인 암, 뇌혈관질환, 심장질환, 당뇨병, 만성 하기도 질환, 간질환, 고혈압으로 보고되었다. 흡연, 음주, 신체활동, 식습관과 같은 생활습관들이 많은 영향을 미치는 것으로 알려져 있으나 정확한 직접적 원인은 아직 밝혀지지 않은 상태이기 때문에 인지 및 치료율, 합병증 관리율 등이 여전히 저조한 상황이다. 특히, 고혈압과 당뇨병은 만성질환 중에서도 유병률이 높고, 지속적인 관리가 중요한 질환으로 잘 알려져 있지만, 국내 통계청 자료에 의하면 매년 유병률은 조금씩 증가하는 양상을 보이고 있다. 조사 결과에 따르면 2014년 기준으로 고혈압 인지율과 치료율은 각각 65.8%와 62.3%로 계속해서 증가하는 추세이지만, 여전히 유병자를 기준으로 하는 고혈압 조절률은 45.7%로 비교적 낮은 수치를 보였으며, 당뇨병의 경우도 마찬가지로 인지율 70.7%, 치료율 63.0%로 높은 수준을 나타냈지만 조절률은 25.3%에 그쳤다.

대부분의 만성질환은 특정한 치료방법이 없어 단일 질병 상태로 유지되기 보다는 시간이 지나면서 합병증을 유발하거나 건강 상태를 악화시켜 심각한 문제를 야기하게 된다. 합병증 발생 후 개별 질병을 관리를 위해 여러 질병을 동시에 치료 시, 장기간에 걸친 치료와 비용에 의해 경제적인 부담과 비효율적 치료로 인해 질병의 악화와 같은 복합적인 문제가 발생할 수 있다(Moussavi et al.,

2007, Barnett et al., 2012). 특히, 최근 들어 고령 인구층의 급격한 증가와 만성질환자의 유병률 향상으로 두 가지 이상의 질환을 가진 다발성 질환(multimorbidity) 환자의 수가 증가하고 있음에도 불구하고, 검사나 진료는 대개 단일 질병을 중심으로 이루어지고 있다(Fortin et al., 2005). 다발성 질환자의 경우 단일 조건을 가진 환자보다 조기에 사망 할 확률이 높고 치료 및 병원 입원 기간이 길기 때문에 면역에 취약한 노인층이나 저소득층에게 많이 나타나며 자칫 생명을 위협하는 심각한 문제로 이어질 수 있다(Vogeli et al., 2007). 해외에서는 이미 지난 몇 년 전부터 많은 의료 기획자와 정부에서 이와 같은 다발성 질환의 문제성을 인지하고 연령 관련 질환에 관한 인구 기반 연구가 시행되고 있다<Table 1.3>. 지난 세기 동안 전세계적으로 만성질환 문제가 타 질환에 비해 압도적으로 증가하고 있으며, 거의 모든 질환 조건이 노화와 밀접한 관련이 있다는 것이 공통된 연구 결과이다. 또한 우리나라 노인 실태 조사(만 65세 이상) 결과, 2011년에 조사된 만성질환자 수는 10년 전에 비해 단일 만성질환 수는 감소하였지만, 두 가지 이상의 질환을 보유한 환자수는 증가하고 있는 추세이다. 하지만 여전히 대부분의 연구는 단일 질병 중심으로 이루어지고 있으며, 노인의 건강 상태와 연령 관련 병리의 복잡성과 이질성을 감안할 때, 모든 연구 및 임상 목적을 효과적으로 수행 할 수 있는 시스템이 구축 되어있지 않은 실정이다(Valderas et al., 2009; Fratiglioni et al., 2010).

따라서 우리 나라에서도 국내 조사 자료를 이용하여 한국인에게 맞는 독자적인 증상 점검 시스템을 구축하고 컴퓨터를 활용한 머신 러닝 기법을 접목시켜 기존의 연구와는 차별화된 결과를 도출해야 할 필요가 있다. 시스템 구축 후, 데이터의 지속적인 관리를 통해 관련 연구자들에게 최신 정보를 제공하고

복합 질환 관련 위험도 예측 분석에 활용 가능한 인프라를 제공하여 국내에서도 활발한 다발성 질환 연구가 이루어져야 할 것이다. 더 나아가 관련 연구자들뿐만 아니라 일반인들에게도 효율적인 정보 검색 및 접근이 가능하도록 하고 사전에 예방 가능한 질병 관리에 유용한 정보를 제공하며 개인의 건강 관련 요인들에 따라 발생 가능한 질병 목록을 함께 제시함으로써 국내 주요 질병의 발병과 유병률을 낮추고 공중보건 향상에도 기여할 수 있을 것이다.

**Table 1.3 Description of the studies concerning multimorbidity**

| Subject of study | No. of subject | Multimorbidity | Reference |
|---|---|---|---|
| Men rural community in Finland, Italy, and the Netherlands | 2285 (65–84 years) | 2 or more co-existing chronic diseases | Menotti et al. (2001) |
| American Indian rural community residents | 1039 (60+ years) | More than 7 diseases increased the risk of 2-year mortality | John et al. (2003) |
| Canada residential | 980 (18+ years) | 2 or more co-existing chronic diseases | Fortin et al. (2005) |
| Australian general population | 17,450 (20+ years) | 3 or more co-existing chronic diseases | Walker (2007) |
| Community dwellers in UK | 9439 (50+ years) | Compared to those with one disease, having 2–3, 4–5, or more than 6 diseases | Kadam et al. (2007) |
| Australian general practitioner patients | 9156 (all ages) | 2 or more co-existing chronic diseases | Britt et al. (2008) |
| Dutch general population | 13584 (all ages) | 2 or more co-existing chronic diseases | Uijen and van de Lisdonk (2008) |
| Swedish living in community and institutions | 1099 (78+ years) | 2 or more co-existing chronic diseases | Marengoni et al. (2008) |
| Spanish living in community | 2192 (20+ years) | 2 or more co-existing chronic diseases | Loza et al. (2009) |
| US Medicare recipients | 2,232,528 (all ages) | 2 or more co-existing chronic diseases | Schneider et al., (2009) |

Ten studies were identified on prevalence of multimorbidity; four included only persons aged 50 years or older and six evaluated people of different ages. Data were available only for a few countries such as Finland, Italy, Netherlands, Sweden, Australia, Canada, Spain, United Kingdom and US.

## 1.5 연구의 목적

혈압이나 혈당의 조절은 지속적인 관리가 필요하기 때문에 복합적인 다발성 질환으로 발생할 수 있는 각종 합병증 위험도를 파악하고 평가하는 것이 중요하다. 또한, 특정 질환을 가지고 있는 환자들 대부분은 여러 가지 질병을 동시에 가지고 있는 경우가 많아 이들 동반 질환에 대한 포괄적인 접근이 필수적이다. 하지만 현재 국내 주요 발병 질환 관리 체계는 과거 급성기 질환을 중심으로 만들어진 것으로, 질환에 대한 예방적 관리 및 자가 관리가 점차 중요해지고 있는 현재의 역학적 변화를 반영하고 있지 못하고 있는 실정이다. 질환의 효과적인 관리를 위해서는 기존의 치료 질병 중심의 접근 방식에서 확장하여 사전에 개인의 행태에 따른 건강 상태를 스스로 점검하고, 발병 가능한 질환에 대한 사전 지식을 습득하여 예방 가능한 시스템 구축이 이루어져야 한다. 특히 만성질환의 경우, 건강 행태와 식습관에 크게 영향을 받는 질병으로 알려져 있기 때문에 국내에서 사용하기에는 앞서 제시한 해외에서의 연구 사례들과 웹사이트의 이용에 많은 한계점이 존재할 수 있다. 따라서 국외 여러 나라의 건강 관리 시스템을 참고하여 우리 국민에 특화된 시스템 구축이 필요하다.

본 연구의 목적은 다년간 축적된 국민 건강영양조사 데이터를 기반으로 우리나라 실정에 맞는 주요 발병 질환 관련 데이터베이스를 구축하고, 구축된 질병 데이터를 활용하여 간편하고 빠르게 각종 질환 관련 증상에 대한 예측도를 산출해내는 시스템을 만드는 것이다. 시스템을 구축한 후, 웹 인터페이스를 통해 이를 개방하여 사용자로 하여금 간단한 기본 정보와 본인의 건강 행태, 증상 및 평소 생활 습관이나 식습관과 같은 개인의 건강 지표가 될 수 있는 요인들에 따른 발병 가능 질환 목록을 제시해 주고, 가능성

있는 질환을 예방하는데 유용한 건강 정보를 함께 제시함으로써, 개인의 질병 유발 가능성을 관리하고, 실질적인 자가 관리를 통해 질병 발병률을 낮추는데 기여할 수 있을 것이다. 또한, 문헌 조사를 통해 얻은 관련 질환 전문적인 정보를 링크와 함께 제공함으로써 해당 질병에 대한 불필요한 검색 과정 없이 질병 지식과 예방 상식과 같은 정보 습득이 가능하도록 할 것이다. 또한, 대용량의 건강 지표 관련 데이터 연구에 활용 가능한 분석 기법을 제시하여 보건학 분야 연구자들이 유용하게 이용할 수 있도록 인프라를 마련하고, 최근 임상 실험 결과 및 질병관련 논문과 문헌 자료들을 적용하여 업데이트 과정을 통해 가능한 최신의 정보를 제공할 수 있도록 하여 연구 동향에 활용될 수 있도록 하고자 한다. 추후 유전적 요인이 크게 작용하는 만성 질환이나 돌연변이에 의한 악성 신생물의 경우, Genbank (http://www.ncbi.nlm.nih.gov/genbank)에서 제공하는 관련 질병 유전자 염기 서열이나 돌연변이 서열과 같은 유전적 데이터를 추가하고, 이와 관련된 질환의 목록 산출 및 돌연변이 패턴 분석을 통해 만성 질환에 대한 다양한 연구가 가능한 인프라를 마련하고자 한다.

# 제 2 장. 연구 방법

## 2.1 데이터 수집 및 가공

본 연구에서는 건강 지표 연관성을 활용한 질병 위험도 예측 연구를 위해 보건정책 수립과 평가에 필요한 통계 산출을 위해 매년 조사를 실시하는 국민건강영양조사(KNHANES) 자료 중 대상자가 직접 작성하는 건강행태조사 설문 자료를 사용하였다. 홈페이지에서 제공하는 원시 자료 중 6기 1차년도(2013), 6기 2차년도(2014), 6기 3차년도(2015), 7기 1차년도(2016)를 기반으로 만성 질환 관련 참고 문헌을 통해 관련 증상들과 관련 지표 및 주요 질환들과 기타 정보들을 세분화하여 정리하였다<Table 2.1, Table 2.2, Table 2.3, Table 2.4>. 그 다음 데이터베이스에 필요한 정보들만을 추출하기 위해서 SAS 포맷으로 제공된 원시 데이터를 먼저 텍스트 파일로 변환시키고, Java 프로그래밍을 통해 MySQL에 저장하여 1차 데이터베이스를 구축하였다. 구축된 1차 데이터베이스에는 제공된 모든 정보를 추출 과정 없이 포함시켰기 때문에 이를 그대로 사용할 시 방대한 데이터양으로 분석 과정에 효율성과 가동성이 떨어지는 단점이 있다. 따라서 이를 보완하기 위해 Java 프로그래밍을 통해 본 연구에서 사용할 데이터만을 추출하는 parsing 과정을 수행하였다.

사전에 조사한 주요 만성질환 건강 지표로 활용 가능한 설문조사 정보들과 실제 의사로부터 과거에 관련 질환 확진 판정을 받은 문항의 응답 정보를 1차적으로 추출하고, 만성

질환의 주요 원인으로 알려진 음주, 흡연과 같은 건강 행태에 대한 정보를 개인 아이디 별로 분류하여 2차적 데이터 추출과정을 통해 본 연구에 사용될 2차 데이터 베이스를 구축하였다. 사용된 총 데이터는 2,134,370개로, 이 중 기본 정보는 성별(31,098), 나이(31,098), 키(29,443), 몸무게(29,474), 허리둘레(29,439), 체질량 지수(29,441), 맥박 규칙성 여부(26,060), 60초 맥박수(1,925), 혈압(26,062)으로 구성되어있다. 증상에는 최근 2주간 몸이 불편했던 경험(57,738), 기침(18,306), 가래(19,138), 혈담(2,914), 흉통(2,969), 호흡곤란(2,920), 체중감소(2,937), 피로감(3,017), 발열(2,902), 야간 발열(2,906), 체중증가(28,870)가 있으며, 음주(282,490)와 흡연(245,129)에 관련된 건강 행태가 있다. 이환 조사 항목은 과거에 의사에게서 확진을 받은 항목들 중 고혈압(57,740), 이상지질혈증(57,740), 뇌졸중(57,740), 심근경색 또는 협심증(55,178), 심근경색 (57,740), 협심증(57,740), 폐결핵(57,740), 천식(57,740), 당뇨(57,740), 갑상선 질환(57,740), 위암(57,740), 간암(57,740), 폐암(57,740), 갑상선 암(57,740), 신부전(57,740), 간경변증(57,740)을 사용하였고, 질병 관련 기타 항목(136,553)와 가중치 값(11,840)으로 이루어져 있다<Table 2.5>. 추출된 대용량의 가공 데이터들을 MySQL에 저장되도록 Java코드를 작성하였으며, 추후 연도별 데이터 분석에 활용될 수 있도록 원시 자료를 각 해당 연도별로 테이블을 만들어 총 4개의 데이터베이스와 통합 분석에 활용 가능하도록 4차년도 데이터를 합친 하나의 데이터베이스, 그리고 일차적으로 본 연구에서 활용될 정보만을 포함시킨 가공 데이터베이스를 구축하였다.

**Table 2.1 Selected risk factors and related symptoms for major chronic disease**

| Category | Code | Type | Description |
|---|---|---|---|
| Personal information | ID | Var(12) | Personal ID |
| | ID_fam | Var(12) | Family ID |
| | Year | Integer | Research year |
| | Sex | Integer | Sex |
| | Age | Integer | Age |
| Measurement information | HE_ht | Decimal(12) | Height |
| | HE_wt | Decimal(12) | Weight |
| | HE_wc | Decimal(12) | Waist size |
| | HE_BMI | Decimal(12) | BMI |
| | HE_rPLS | Integer | Pulse regularity |
| | HE_mPLS | Integer | Pulse per min |
| | HE_nARM | Integer | Blood pressure |
| Subjective symptoms | D_2_1 | Integer | Experience in the past two weeks with discomfort |
| | D_2_wk | Integer | Discomfort for the last 2 weeks |
| | HE_cough1 | Integer | Cough experience for 3 consecutive months or more |
| | HE_cough2 | Integer | Period of cough experience |
| | HE_sput1 | Integer | Phlegm experience for 3 consecutive months or more |
| | HE_sput2 | Integer | Period of phlegm experience |
| | HE_tb1 | Var(225) | Cough manifestation |
| | HE_tb1_d | Var(225) | Cough manifest days |
| | HE_tb2 | Var(225) | Phlegm manifestation |
| | HE_tb2_d | Var(225) | Phlegm manifest days |
| | HE_tb3 | Var(225) | Blood phlegm manifestation |
| | HE_tb3_d | Var(225) | Blood phlegm manifest days |
| | HE_tb4 | Var(225) | Chest pain manifestation |
| | HE_tb4_d | Var(225) | Chest pain manifest days |
| | HE_tb5 | Var(225) | Dyspnea manifestation |
| | HE_tb5_d | Var(225) | Dyspnea manifest days |
| | HE_tb6 | Var(225) | Notable weight loss |
| | HE_tb6_d | Var(225) | Weight loss days |
| | HE_tb7 | Var(225) | Tiredness |
| | HE_tb7_d | Var(225) | Tiredness days |
| | HE_tb8 | Var(225) | Fever manifestation |
| | HE_tb8_d | Var(225) | Fever manifest days |
| | HE_tb9 | Var(225) | Night fever manifestation |
| | HE_tb9_d | Var(225) | Night fever manifest days |

**Table 2.2 Selected health behavior that can have an indirect effect**

| Category | Code | Type | Description |
|---|---|---|---|
| Health behavior (drinking) | BD1 | Integer | Drinking experience (12 years or older) |
| | BD2 | Integer | Start drinking age (12 years or older) |
| | BD1_11 | Integer | 1-year frequency of drinking |
| | BD2_1 | Integer | The amount of alcohol consumed at once |
| | BD2_31 | Integer | Frequency of binge |
| | BD2_32 | Integer | (Female) Frequency of binge |
| | BD4 | Integer | Frequency of drinking out of control |
| | BD5 | Integer | Frequency of drinking disrupted daily lives |
| | BD6 | Integer | Frequency of drinking for hangover |
| | BD7_1 | Integer | Frequency of regret after drinking |
| | BD7_2 | Integer | Frequency of oblivion after drinking |
| | BD7_3 | Integer | Wound after drinking |
| | dr_month | Integer | Monthly drinking rate |
| Health behavior (smoking) | BS1_1 | Integer | Lifetime smoking |
| | BS3_1 | Integer | Currently smoking |
| | BS3_2 | Integer | Average smoke per day |
| | BS6_2 | Integer | (Past smoker) Period of smoking |
| | BS6_3 | Integer | (Past smoker) Amount of smoking |
| | BS8_2 | Integer | Indirect exposure to secondhand smoke in the workplace |
| | BS9_1 | Integer | Household smoker existence |
| | BS9_2 | Integer | Indirect exposure to secondhand smoke in household |
| | BS13 | Integer | Indirect exposure to secondhand smoke in public institutions |

In tables 2.1 and 2.2, the elements that can be used as health indexes among the survey data are extracted and listed through various reference documents and specialized books. When constructing the database, the code name is used to indicate the name of each column and the data type is indicated together.

**Table 2.3 Selected major chronic disease**

| Category | Code | Type | Description |
|---|---|---|---|
| Chronic diseases | DI1_dg | Integer | Diagnosed hypertension |
| | DI1_pr | Integer | Hypertension present condition |
| | DI2_dg | Integer | Diagnosed dyslipidemia |
| | DI2_pr | Integer | Dyslipidemia present condition |
| | DI3_dg | Integer | Diagnosed stroke |
| | DI3_pr | Integer | Stroke present condition |
| | DI4_dg | Integer | Diagnosed myocardial infarction/angina pectoris |
| | DI4_pr | Integer | Myocardial infarction/angina pectoris present condition |
| | DI5_dg | Integer | Diagnosed myocardial infarction |
| | DI5_pr | Integer | Myocardial infarction present condition |
| | DI6_dg | Integer | Diagnosed angina pectoris |
| | DI6_pr | Integer | Angina pectoris present condition |
| | DJ2_dg | Integer | Diagnosed tuberculosis |
| | DJ2_pr | Integer | Tuberculosis present condition |
| | DJ4_dg | Integer | Diagnosed asthma |
| | DJ4_pr | Integer | Asthma present condition |
| | DE1_dg | Integer | Diagnosed diabetes mellitus |
| | DE1_pr | Integer | Diabetes mellitus present condition |
| | DE2_dg | Integer | Diagnosed thyroid disease |
| | DE2_pr | Integer | Thyroid disease present condition |
| | DC1_dg | Integer | Diagnosed stomach cancer |
| | DC1_pr | Integer | Stomach cancer present condition |
| | DC2_dg | Integer | Diagnosed liver cancer |
| | DC2_pr | Integer | Liver cancer present condition |
| | DC6_dg | Integer | Diagnosed lung cancer |
| | DC6_pr | Integer | Lung cancer present condition |
| | DC7_dg | Integer | Diagnosed thyroid cancer |
| | DC7_pr | Integer | Thyroid cancer present condition |
| | DN1_dg | Integer | Diagnosed renal failure |
| | DN1_pr | Integer | Renal failure present condition |
| | DK4_dg | Integer | Diagnosed liver cirrhosis |
| | DK4_pr | Integer | Liver cirrhosis present condition |

Among the findings of the questionnaire survey, it is a list of chronic diseases which is the main cause of death in Korea.

**Table 2.4 Other related factors and weight**

| Category | Code | Type | Description |
|---|---|---|---|
| Other related information | BP8 | Integer | Average sleep time per day |
| | BP1 | Integer | About the usual stress |
| | BP5 | Integer | Continuous depression for more than 2 weeks |
| | BO3_14 | Integer | Weight control method: Absent |
| | BO3_05 | Integer | Weight control method: weight loss drug |
| Given weight | wt_pft | Double | Pulmonary function test weight |
| | wt_ex1 | Var(225) | Chest X-ray inspection weight |
| | wt_ex1pf | Var(45) | Chest X-ray inspection and Pulmonary function test weight |

Among the responses, references were used to extract other surrogates that could affect chronic diseases. In addition, we used the weighted value for chest X-ray examination because it is given.

**Table 2.5 Type and number of data used for system construction**

| Category | Data (Number of data) |
|---|---|
| Personal information | Sex (31,098), Age (31,098), Height (29,443), Weight (29,474), Waist size (29,439), BMI (29,441), Pulse regularity(26,060), Pulse per min(1,925), Blood pressure(26,062) |
| Symptoms | Discomfort (57,738), Cough (18,306), Phlegm (19,138), Blood phlegm (2,914), Chest pain (2,969), Dyspnea (2,920), Notable weight loss (2,937), Tiredness (3,017), Fever (2,902), Night fever (2,906), Weight gain (28,870) |
| Health behavior | Drinking (282,490), Smoking(245,129) |
| Disease | Hypertension (57,740), Dyslipidemia (57,740), Stroke (57,740), Myocardial infarction/angina pectoris (55,178), Myocardial infarction (57,740), Angina pectoris (57,740), Tuberculosis (57,740), Asthma (57,740), Diabetes (57,740), Thyroid (57,740), Stomach cancer (57,740), Liver cancer (57,740), Lung cancer (57,740), Thyroid cancer (57,740), Renal failure (57,740), Liver cirrhosis (57,740) |
| Others | Other related information (136,553), Weight (11,840) |

This table shows the type and number of data used in this study. Some of data shows only specific year, and health behavior is a combination of details.

## 2.2 시스템 개발 환경 및 서버 구축

웹 기반의 데이터베이스 구축을 위해 Intel Xeon E5-240 v2 2.40GHz CPU, 8GB RDIMM (1600MT/s) memory, 300GB 15K RPM / 6Gbps HDD (OS), 3TB 7.2K RPM Near Line 6Gbps HDD (DATA)사양의 서버를 사용하였다. CPU는 8C AMD Opteron-6128 2.0GHz x 1로 8Core의 사양을 가지고 있는 것을 이용하였으며, 사용된 서버의 운영체제는 Linux 기반의 v2.6.18이다. 제공된 데이터의 포멧은 SAS타입으로 이를 가공하기 위해 파일 타입을 변화하는 작업을 수행하였다. 가공된 데이터는 eclipse 프로그램을 통해 Java 프로그래밍 언어를 이용하여 관계형 데이터베이스 관리 시스템인 MySQL v5.5에 저장하였다. 데이터베이스에 데이터를 업데이트 할 수 있는 환경을 제공하였으며, 이를 기반으로 Java로 코딩 된 프로그램 내에서 SQL 쿼리문을 수행할 수 있도록 데이터베이스와 연결하는 작업을 수행하는 JDBC (Java Database Connectivity)를 이용하여 MySQL과 연동하였다. 웹 기반의 데이터베이스를 구현하기 위해 웹 페이지를 생성하고 웹 브라우저에서 시각화해주는 언어인 HTML (Hyper Text Markup Language), JSP (Java Server Page), servlet, JavaScript를 이용하였다. JSP와 servlet은 Java를 기반으로 하는 웹 애플리케이션 프로그래밍 기술로, 서블릿 클래스를 이용하여 패키지에 속하는 인터페이스들을 구현하였다. Java프로그램을 컴파일 하고 실행시키기 위해서는 JDK (Java development kit) v1.8.0을 사용하였다. 사용된 알고리즘은 JCUDA라이브러리 검색을 통해 구현되었으며, Weka를 통해 분석 과정이 이루어졌다. 마지막으로 생성된 웹 인터페이스 서비스는 웹 컨테이너인 Tomcat v 8.5.27을 이용하여 서버에 담긴 정보들을 연동하여 최종적으로 구현하였다.

## 2.3 데이터베이스 구현

구현한 데이터베이스의 궁극적인 목적은 관련 연구자들에게 빠르고 편리한 통합적 만성질환의 건강 지표가 될 수 있는 증상이나 건강 행태와 같은 데이터를 제공하고 이를 데이터베이스화 시켜 머신 러닝 알고리즘을 활용하여 질병 요인을 분석하고 위험도 예측 연구에 도움이 되는 플랫폼을 제공하는 것이다. 본 연구에서는 4차년도에 걸친 대용량의 데이터들을 1차적으로 분리하여 '2013', '2014', '2015', '2016' 4개의 raw data table을 생성하였다. 그 다음 Java 프로그래밍을 통해 기본 정보와 건강 검진 정보들을 변수 명을 이용하여 추출한 다음 이를 텍스트 파일 형식으로 저장하였다. Java코드를 작성하여 텍스트 파일의 정보를 다시 MySQL에 각각 '2013_1', '2014_1', '2015_1', '2016_1'의 데이터베이스를 구축하였다.

각 컬럼은 원시 자료에서 제공하는 기본 변수를 사용하였으며, 개별 아이디, 가족 아이디의 경우 필드에 'varchar' 데이터 타입으로 저장하고, 수치 정보를 분석에 이용하고자 하는 데이터의 경우 'integral' 타입으로 저장하였다. 아이디, 조사 연도, 성별, 나이, 신장, 체중, 허리둘레, BMI, 혈압과 같은 기본 정보는 총 12개의 컬럼으로 구성되어 있으며 기침, 가래, 흉통, 호흡곤란, 피로감, 체중감소, 발열, 혈담 등과 같은 증상 관련 정보의 경우 24개의 컬럼으로 각 필드 타입은 'integral'과 'varchar'로 저장하였다. 건강 행태에 관한 설문 자료 중 음주, 흡연에 관한 정보를 사용하였으며, 음주 경험의 유무, 음주 시작 연령, 음주 빈도수 월간 음주 율 등의 13개 컬럼과 흡연 여부, 현재 흡연 여부, 흡연 기간, 과거 및 현재 평균 흡연량 등의 9개 정보를 'integral'과 'varchar' 타입으로 저장하여 총 47개의 증상 및 행동 데이터로 저장하였다. 만성질환의 경우 과거나 현재에

의사에게서 진단을 받은 경우와 현재 유병의 여부 응답을 이용하여 총 63가지의 컬럼을 생성하여 저장하였다. 마지막으로 평균 수명 시간, 체중 조절 방법, 스트레스 인지 정도, 우울 감과 같은 기타 특이적인 사항들과 원시자료에서 제공된 가중치 값들 또한 ‘integral’, ‘varchar’, ‘decimal’의 형태로 필드에 입력해주었다<Figure 2.1>. 전체는 총 99개의 컬럼으로 구성되어있으며, 2013년도 8,019개, 2014년도 7,551개, 2015년도 7,381개, 2016년도 8,151개로 총 31,102개의 열로 이루어져있다. 통합 데이터베이스 가공 과정에서 간혹 특정 연도에만 조사된 내용의 데이터가 존재하는 경우 이를 포함하지 않는 데이터에도 조사 코드를 포함시키고 ‘null’ 값으로 처리하여 해당 데이터 사용시 오류가 나지 않도록 하였다. 마지막으로 구축될 시스템을 이용하여 분석 시 각 명령어를 수행하는데 있어서 목차의 역할을 해 줄 수 있도록 인덱스 키를 설정하여 이를 이용한 검색 또한 가능하도록 해 주었다. 편리한 정보의 이용을 위해 검색 기능을 추가하여 사용자가 원하는 만성질환의 정보를 쉽게 찾아 볼 수 있으며, 변수 코드와 설명에 관한 정보를 테이블로 시각화하여 웹 상에서 확인이 가능하도록 하여 각 연도 별 데이터 중 필요한 변수들을 선택하여 비교하여 볼 수 있도록 하였다.

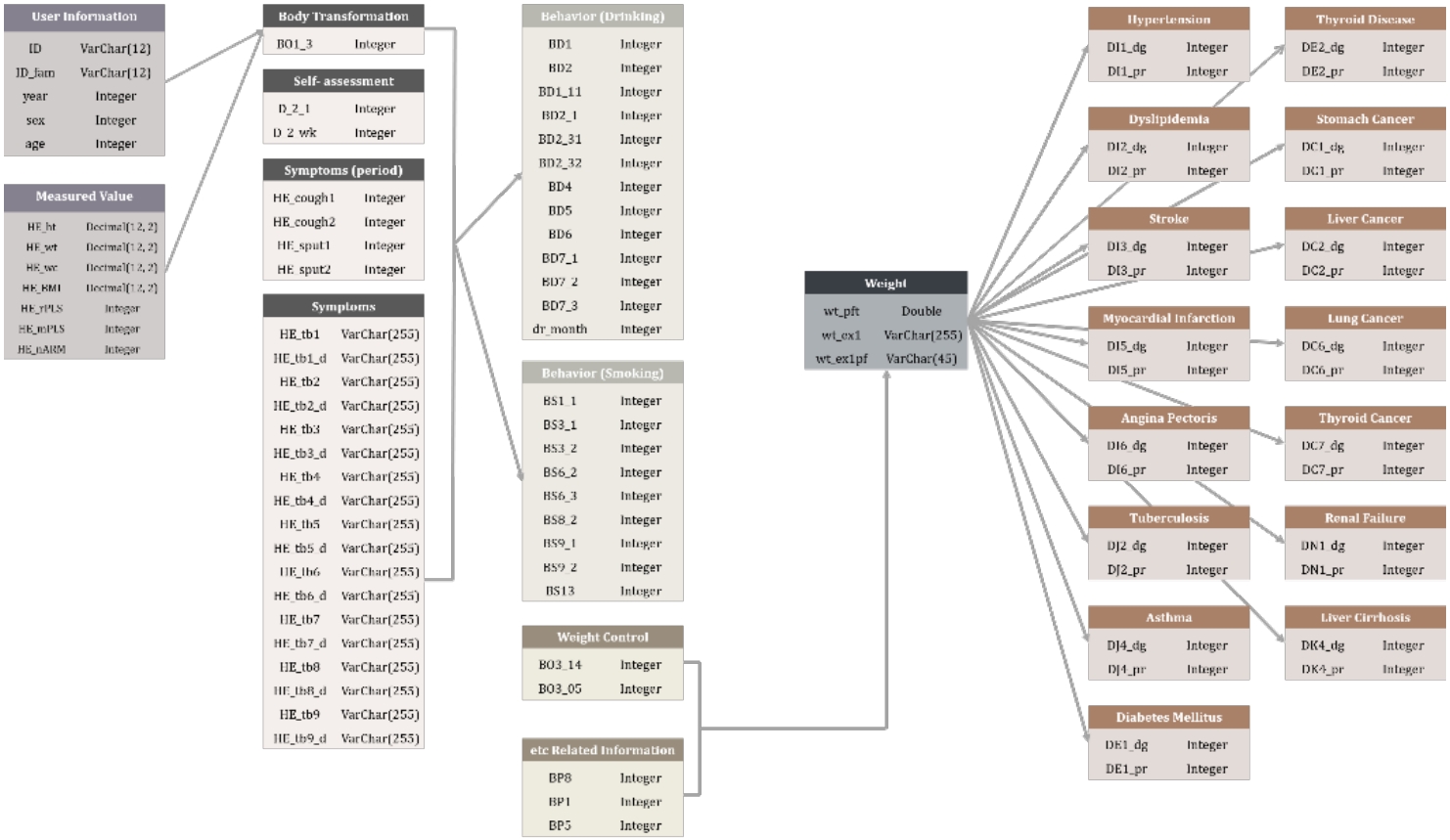


**Figure 2.1 Database schema.** The diagram shows only the major tables in the database and which files should be linked with which other fields when creating queries. After receiving the basic information query from the user, it passes the query value related to the symptoms. Then, the value according to the health behavior such as drinking and smoking is received, and it is handed to the query value as the diagnosis result according to the weight value. The information about the disease risk and the possible disease according to the symptom and the body measurement is shown.

## 2.4 머신 러닝을 활용한 분석

### 2.4.1 머신 러닝과 라이브러리

본 연구에서 사용되는 Java 는 현재 이용 가능한 다수의 머신 러닝 라이브러리가 있어 기존의 Java 애플리케이션과 편리하게 인터페이스로 연결 할 수 있고, 머신 러닝의 성능 또한 극대화 시킬 수 있다는 장점이 있다. 머신 러닝 오픈 소스 소프트웨어 커뮤니티(http://www.mloss.org/software/)에는 현재 약 80 여개의 Java 기반 오픈소스 머신 러닝 프로젝트가 있으며, 각 대학교의 서버나 깃 허브(https://github.com/), 비트버켓(https://bitbucket.org/) 등에도 많은 프로젝트가 존재한다(Gvero and Kuncak, 2015). 머신 러닝 애플리케이션 아키텍처를 구성하는데 있어서 프로젝트의 목적, 조건에 따라 어떤 머신 러닝 라이브러리를 선택하는가 는 중요한 문제이다. 현재 전 세계적으로 많이 쓰이는 주요 Java 라이브러리는 다음과 같다<Table 2.6>.

Weka (Waikato Environment for Knowledge Analysis, https://www.cs.waikato.ac.nz/ml/weka/)는 뉴질랜드 와이카토 대학에서 개발한 머신 러닝 라이브러리로 현재 가장 널리 알려진 Java 라이브러리이다. Weka 는 범용 라이브러리로 분류, 회귀분석, 클러스터링 등 머신 러닝과 관련된 다양한 임무를 처리하며, 그래픽 사용자 인터페이스와 커맨드 라인 인터페이스, Java API 를 제공한다. Weka 는 자체적인 기본 데이터 포맷인 ARFF 를 지원하는데, ARFF 는 속성-데이터(attribute-data) 쌍을 특징으로 한다(Garner, 1995). 총 두개의 부분으로 구성되어있는데, 첫번째 부분은 헤더를 포함하며 모든 속성과 nominal, numeric, data, string 등의 데이터 타입으로 정의된다. 두번째 부분은 실제 데이터를 포함하고 있으며, 각각의

라인은 하나의 인스턴스에 대응한다. 헤더의 마지막 속성은 암묵적으로 타겟 변수로 간주되며, 데이터가 누락된 경우 물음표를 사용한다(Frank et al., 2009). Java-ML (http://java-ml.sourceforge.net/)은 동일한 유형의 알고리즘에 공통 인터페이스를 적용한 머신 러닝 알고리즘의 집합체로 데이터 전처리, 기능 선택, 분류, 클러스터링 등의 알고리즘이 포함되어 있다(Abeel et al., 2009). Weka 와 마찬가지로 범용 머신 러닝 라이브러리이며 좀 더 일관된 인터페이스를 제공하고 다른 패키지에는 없는 최신 알고리즘과 유사성 측정 및 기능 선택 기술인 다이나믹 타임 워핑(dynamic time warping), 랜덤 포레스트 평가(random forest attribute evaluation) 등을 제공한다(Kaluza, 2016). 또한, 하나의 라인에 하나의 데이터 샘플을 포함하고 세부 속성으로 쉼표, 세미콜론, 탭 등으로 구분되어 있으면 어떤 타입의 파일도 지원 가능하다(Abeel et al., 2009). 또 다른 라이브러리 Mahout (https://mahout.apache.org/)은 확장성 높은 데이터 분산 처리 구조를 지니고 있으며, 서버 클러스터를 이용한 병렬식 분산 처리 알고리즘을 대규모 데이터 세트를 처리하고 생성할 수 있는 맵리듀스(MapReduce) 패러다임이 적용 되어있다(Lyubimov and Palumbo, 2016). Spark (https://spark.apache.org/) Mahout 과 마찬가지로 대규모 데이터 처리 플랫폼이지만 맵리듀스 대신 인 메모리 캐시를 이용해서 데이터 중 필요 부분을 추출, 처리하며, 쿼리를 반복적으로 전송한다. 다양한 모듈로 구성되며, 그래프 처리를 위한 Graph X, 실시간 데이터 스트림 처리를 위한 spark streaming, 그리고 분류, 회귀분석, 협력 식 필터링, 클러스터링, 차원 감소, 최적화 등 머신 러닝 라이브러리 기능 제공을 위한 MLlib 등이 있다(Zaharia et al., 2016). DL4J (Deeplearning4java, https://deeplearning4j.org/) 는 Java 로 작성된 딥 러닝 라이브러리로 피드포워드(feed-forward) 신경망, RBM, 나선형 신경망(convolutional neural nets), 딥 빌리프 네트워크(deep

belief network), 오토 인코더(autoencoders) 등 다양한 신경망 알고리즘을 제공한다(Kaluza, 2016). 마지막으로 MALLET (Machine learning for language toolkit, http://mallet.cs.umass.edu/)는 자연어 처리 알고리즘과 유틸리티를 제공하는 대규모 라이브러리로 문서 분류, 문서 클러스터링, 정보 추출, 주제 모델링 등과 같은 다양한 임무를 수행한다. 나이브 베이즈(naive Bayes), HMM, Latent Dirichlet 토픽 모델, 로지스틱 회귀분석, 조건적 임의 필드(conditional random fields)등 다양한 알고리즘을 사용할 수 있도록 Java API 와 함께 커맨드 라인 입력 방식을 제공한다(McCallum, 2002).

### 2.4.2 신경망 구현 및 적용

머신 러닝을 수행하기에 앞서 먼저 문제에 적합한 머신 러닝 방법과 사용할 특징을 결정하고, 사용할 모델의 파라미터 설정을 결정하여 실제 문제를 처리할 분류기를 만드는 전처리(preprocessing) 작업이 필요하다<Figure 2.2>. 주제 분석(subject analysis)은 적절한 분류기(classifier)를 만드는 과정으로 원시 자료로부터 어떤 특징을 만들어낼 것인가를 결정하는 작업으로, 원시 데이터 중 6기 1차년도, 2차년도, 3차년도 데이터를 학습용 데이터로 7기 1차년도 데이터를 검증 용 데이터로 분리한 다음, 학습용 데이터 셋에서 특징을 추출하였다. 일반적으로 특징을 추출하는 방법은 사람의 경험이나 직관에 의존하는 경우가 많으며, 모델의 이론과 공식이 각 모델별로 상이하므로 최상의 결과를 얻기 위해서는 많은 시간과 노력이 필요로 한다. 따라서 본 연구에서는 이러한 과정을 효율적으로 처리하기 위해 분류 알고리즘과 클러스터링을 통해 데이터들 사이의 관계를 파악하고 분석하는 작업을 실행하였다. 먼저 데이터를 Weka에 불러오기 위해서는 arff 포맷의 파일로 변환시켜주는 작업이 필요하다. 따라서 4차년도 데이터가 모두

들어있는 텍스트 파일을 arff 파일로 변화시켜주는 Java 코드를 작성하여 가공 데이터 포맷을 변경하였다<Figure 2.4>. 학습 모델에 사용할 관련 속성을 지닌 하위 데이터 세트를 선택하기 위해 Weka의 Attributeselection 객체를 통해 전처리 주요 과정인 특징 선택(feature selection)을 수행하였다. 특정 속성이 어느 정도 정보성이 높은지 평가하기 위해 'evaluator' 파라미터와 'evaluator'가 매긴 점수에 따라 속성에 순위를 부여하는 'ranker' 파라미터를 추가하였다. 분류 학습 모델을 만들기 위해 우선 의사결정 트리(decision tree) 객체를 정보성이 가장 높은 속성을 남기고, 정보성이 낮은 속성을 제거하여 좀 더 신속하고 정확한 예측이 가능한 러닝 모델을 구현하였다. Weka에 내장된 기능인 TreeVisualizer를 이용하여 트리의 구조를 확인하고 루트 노드(root node)부터 속성의 일치 여부를 따라 가지를 타고 내려오면서 최종 리프 노드(leaf node)까지 확인하는 작업을 수행하였다. MLP을 구현하기위해 먼저 'MLP'패키지를 만들고 신경망 알고리즘의 기본 절차로 구성된 'MultiLayerPerceptrons.java' 클래스와 역 전파의 실제 코드가 들어있는 'HiddenLayer.java' 클래스를 생성하였다. 출력층에는 다 범주 로지스틱 회귀를 이용한 'LogisticRegression.java' 클래스를 사용하였다. 'ActivationFunction.java'에는 시그모이드 함수의 기울기와 쌍곡 탄젠트(hyperbolic tangent, tanh)를 추가하였으며, 'RandonGenerator.java'에는 정규분포를 갖는 난수(random number)를 생성하는 메소드를 작성하여 은닉층의 가중치를 무작위로 초기화할 수 있도록 하였다. MLP의 파라미터는 은닉층 HiddenLayer와 LogisticRegression의 가중치 W와 바이어스 b로 나타내었으며, W는 유닛 수에 맞춰 무작위로 초기화 될 수 있도록 코드를 작성하였다. 이때 초기값이 잘 분포되어 있지 않으면 지역 최소값 문제가 빈번하게 발생할 수 있으므로 임의 난수 시드(random seed)를

사용하여 실행시켰다. 더 효율적으로 데이터마이닝 부분에 집중하기 위해 라이브러리를 이용한 알고리즘으로 추가 분석을 실행하였다. 현재 자바 기반으로 실무에 사용 가능한 라이브러리인 DL4J에서는 내부적으로 ND4J (N-Dimensional Array for Java) 라이브러리를 이용하고 있다(http://github.com/deeplearning4j/nd4j). ND4J는 다양한 n차원 행렬 객체를 다룰 수 있게 해주는 과학 기술용 라이브러리로 GPU컴퓨팅도 지원 가능하다. 라이브러리를 사용하기 위해 먼저 Maven 프로젝트를 생성한 다음 패키지 명을 DLWJ.ND4J로 지정해주고, 클래스 명을 Perceptron.java로 설정해주었다. CVS파일로 변환시켜준 데이터 베이스를 사용하기 위해 CVS리더를 초기화 해주었으며, CVSRecodReader클래스를 이용하여 데이터를 로드 하였다. Weka를 통해 분석한 방법과 달리 .splitTestAndTrain () 메소드를 사용하여 데이터를 섞어준 다음 데이터를 학습용과 검증용으로 나누어 저장해 주고 본격적으로 모델을 구축하기 위해 시작 부분에 모델 구성을 정의해 주어 망을 설정해 주었다. 이때 제어(regularization) 코드를 추가해 모델이 과적합되는 것을 방지하고 좀 더 일반화된 결과를 출력할 수 있도록 하였다. 명령어 .useDropConnect ()를 사용하여 드롭 아웃이 가능하도록 하였으며, .list ()에 입력층을 제외한 전체 layer수를 정의해주었다. 은닉층의 구성을 설정하기 위해 RBM층을 정의해주고, 모델의 연속 값을 다룰 수 있도록 RBM.VisibleUnit.GAUSSIAN을 사용하였다. 또한, 학습률을 최적화하기 위해 Updater.ADAGRAD를 사용하였다. 모델 구축을 완료한 다음 망을 학습 시키기 위해 프로세스 log코드와 model.fit (train)을 사용하였으며, 평가를 위한 변수들을 설정해주었다. 마지막으로 반복 횟수를 늘리거나 시드 값을 변경하면서 파라미터 값을 조정해주었다.

### 2.4.3 클러스터링을 활용한 분석

감독 분류(supervised classifier) 알고리즘과 비교했을 때, 클러스터링의 목표는 라벨이 붙지 않은 데이터 세트를 각 데이터 고유의 특성으로 구분되는 그룹별로 묶는 것이다. 클러스터링은 동종 그룹화, 유용하면서도 적절한 그룹으로 묶기 또는 이상점(outliers)등 분류의 기준에서 벗어나는 비정상적인 데이터를 찾아내는데 사용된다. 본 연구에서는 31,102개의 인스턴스가 지닌 나이, 성별, 신장, 체중, BMI 등과 같은 기본정보 12개, 증상 및 행동에 관한 변수 47개, 기타 정보 5개를 포함하여 총 64개의 속성으로 구성하였다. 클러스터 모델을 만드는 방법은 분류 알고리즘 모델을 만드는 방법과 유사하며, 알고리즘은 weka.cluster패키지를 이용하여 구현하였다. 분석 과정에서는 예측 최대화(expectation maximization, EM) 클러스터링 기법을 적용하여 설문 조사 응답자들을 공통 속성을 지닌 그룹으로 분류하는 작업을 진행하였다. EM 기법은 개별 인스턴스를 특정 클러스터에 속할 확률 분포에 따라 할당해 주고, 각 클래스의 확률 분포에 대한 파라미터 벡터 값을 다시 측정해주었다. 그 다음 파라미터가 수렴되거나 반복 횟수가 최대에 이를 때까지 반복 실행해주었다. 도출된 클러스터의 숫자는 교차 평가를 통해 수동 또는 자동으로 EM을 설정하는데 사용하였다. 데이터세트에서 클러스터의 숫자를 결정할 수 있는 또 다른 방법인 엘보우 기법(elbow method)을 활용하여 특정 숫자의 클러스터가 존재할 때의 분산 비율을 확인하였다. 클러스터의 증가를 통해 추가적인 정보를 얻을 수 없는 수준까지 즉, 분산의 증가가 거의 없는 수준까지 클러스터를 늘려 분류의 정확도를 높였다.

**Table 2.6 Machin learning library for Java**

| Package | Usage | License | Architecture | Algorithms |
|---|---|---|---|---|
| Weka | versatile | GNU, GPL | single machine | Decision trees, naïve Bayes, neural network, random forest, AdaBoost, hierarchical clustering |
| Java-ML | versatile | GNU, GPL | single machine | k-means clustering, self-organizing map, Markov chain clustering, Cobweb, random forest, decision trees, bagging, distance measures |
| Mahout | classification, clustering | Apache 2.0 | distributed processing, single machine | Logistic regression, naïve Bayes, random forest, HMM, multilayer perceptron, k-means clustering |
| Spark | versatile | Apache 2.0 | distributed processing | SVM, logistic regression, decision trees, naïve Bayes, k-means clustering, linear least squares, LASSO, ridge regression |
| DL4J | deep learning | Apache 2.0 | distributed processing, single machine | RBM, deep belief network, deep autoencoders, recursive neural tensor network, convolutional neural network |
| MALLET | text mining | Common public license 1.0 | single machine | Naïve Bayes, decision trees, maximum entropy, hidden Markov models, conditional random fields |

The main libraries popularly used for Java user. There are many libraries to solve the characteristic problems of each area.

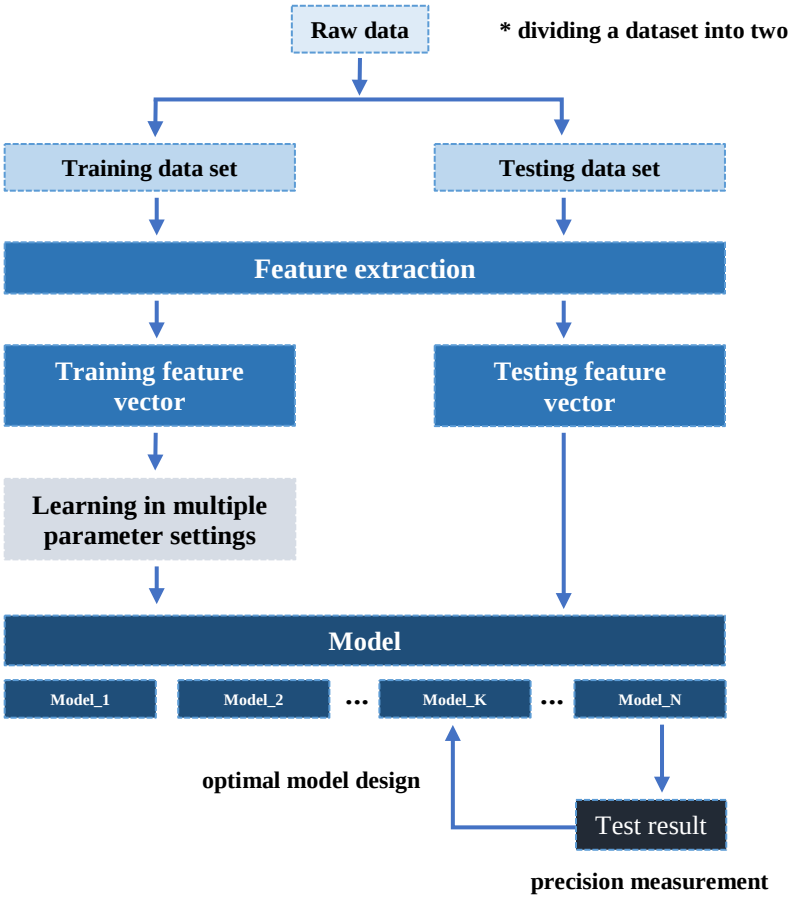


**Figure 2.2 Machine learning process.** The following diagram provides an overview of the overall learning process for machine learning: updating and adjusting the model parameters in the learning phase, and evaluating the value of the model in the testing phase. Repeat the learning, testing, and learning processes until find a suitable value.

## 2.5 알고리즘 성능 평가

예측 알고리즘의 학습은 너무 복잡하거나 너무 단순하지 않고 적절한 수준을 유지하는 것이 중요하다. 복잡성이 너무 낮은 모델(leftmost models)의 경우 최빈값 또는 평균값에 준하는 예측치를 도출 할 수 있으며, 복잡성이 지나치게 높은 모델(rightmost models)은 오직 학습 표본에서 도출된 값을 기준으로 예측할 수 밖에 없다는 문제점을 가지고 있다(Bayan et al., 2008). 지나치게 견고한 모델은 복잡한 패턴을 파악할 수 없는 반면, 지나치게 유연한 모델은 학습 데이터에 포함된 노이즈까지 반영할 수 있다. 머신 러닝 커뮤니티의 중론에 따르면, 일반화 오류의 정도를 측정하기 위해서는 먼저 학습용 데이터와 검증용 데이터를 70:30 비율로 나누는 것이 적절하다(Chan and Stolfo, 1998). 학습용 데이터로 예측 알고리즘을 훈련 시키고 검증용 데이터로 예측을 시킨 다음, 예측 치와 실제 값 사이의 차이를 확인하여 실존하는 일반화 오류를 측정 할 수 있다. 오류의 측정은 다음 두 가지 가정에 근거한다. 검증 데이터 세트는 모집단으로부터 편향됨이 없이 추출되었으며, 실제 사용하게 될 새로운 데이터는 기존에 사용해온 훈련 데이터, 검증 데이터와 같은 분산을 따른다(Bloedorn et al., 1997). 첫번째 가정의 경우, 교차 평가(cross-validation)와 사례를 통해 확인이 가능하며, 모집단에서 별도의 검증 데이터 세트를 분리하기 어려울 정도로 데이터의 양이 충분하지 못하면 학습 알고리즘이 제대로 성능을 발휘하기가 어려워 질 수 있다. 교차 평가는 데이터 세트를 거의 동일한 크기의 n개의 세트로 분할한 다음, 이 중 1개를 검증용으로 나머지 n-1개를 학습용 데이터로 사용하고 전체 데이터세트에 대해 이와 같은 방식을 반복적으로 적용하면서 나타나는 오류에 대한 평균을 구한다. 이러한 방법으로 모든 데이터를 이용하여 학습 및 검증을 진행할 수 있으며, 동시에 특정

모델에서 같은 데이터를 가지고 학습과 검증하는 일도 피할 수 있다. 리브 원 아웃(leave-one-out)은 교차 평가의 극단적인 사례 중 하나로 학습-검증 데이터 조합의 수(fold)는 표본 수와 같다(Kearns and Ron, 1999; Cawley and Talbot, 2003). 표본 중 하나의 데이터 세트에 대해서만 학습하고 그 외의 모든 데이터 세트는 검증용으로 사용되며, 전체 표본을 대상으로 시행 시 각각의 표본을 정확히 한 번씩 평가할 수 있게 된다. 계층화(stratification)는 각각의 폴드가 대략 비슷한 클래스 값 비율을 보이는 경우 표본의 하위 데이터 세트를 선택하는 방식이다(Diamantidis et al., 2000). 특정 클래스가 연속적인 경우 폴드의 평균값을 선택하면 다른 폴드에서도 거의 비슷하게 나타나는 성향을 이용하는 방식이다. 계층화는 교차 평가 방식과 함께 사용할 수 있으며, 별도의 학습-검증 데이터 세트에도 적용할 수 있다.

본 연구의 분석과정에서 사용된 MLP의 정확도, 정밀도, 재현률 값을 백분율로 출력하기 위해 학습된 망으로 입력 데이터를 활성화 시키고, 출력 확률을 label에 넣어 predict 메소드를 실행하였다<Figure 2.3>. 또한, DL4J 라이브러리를 이용한 분석 방법의 평가를 위해 변수를 설정해 주고, eval.eval () 메소드를 사용하여 특징 매트릭스(feature matrix)의 값들을 구해주었다. 이 과정을 통해 accuracy, precision, recall, F1 score 값을 얻을 수 있었다. 클러스터 모델의 평가 방법은 일차적으로 클러스터링의 알고리즘 품질 평가를 시행하였다. 각각의 클러스터가 지닌 데이터의 일관성을 분석하는 로그가능도(loglikelihood) 방식으로 측정하였다. 데이터 세트는 여러 개의 폴드로 나누고 클러스터링 알고리즘은 각 폴더 단위로 실행되었으며, 파라미터로 적합하지 않은 유사한 데이터에 높은 확률을 부여할 경우 해당 데이터 구조의 윤곽을 파악하는데 활용하였다. 이를 위해 Weka의 ClusterEvaluation 클래스를 이용하여 클러스터링 성능 평가를 실행하였다.

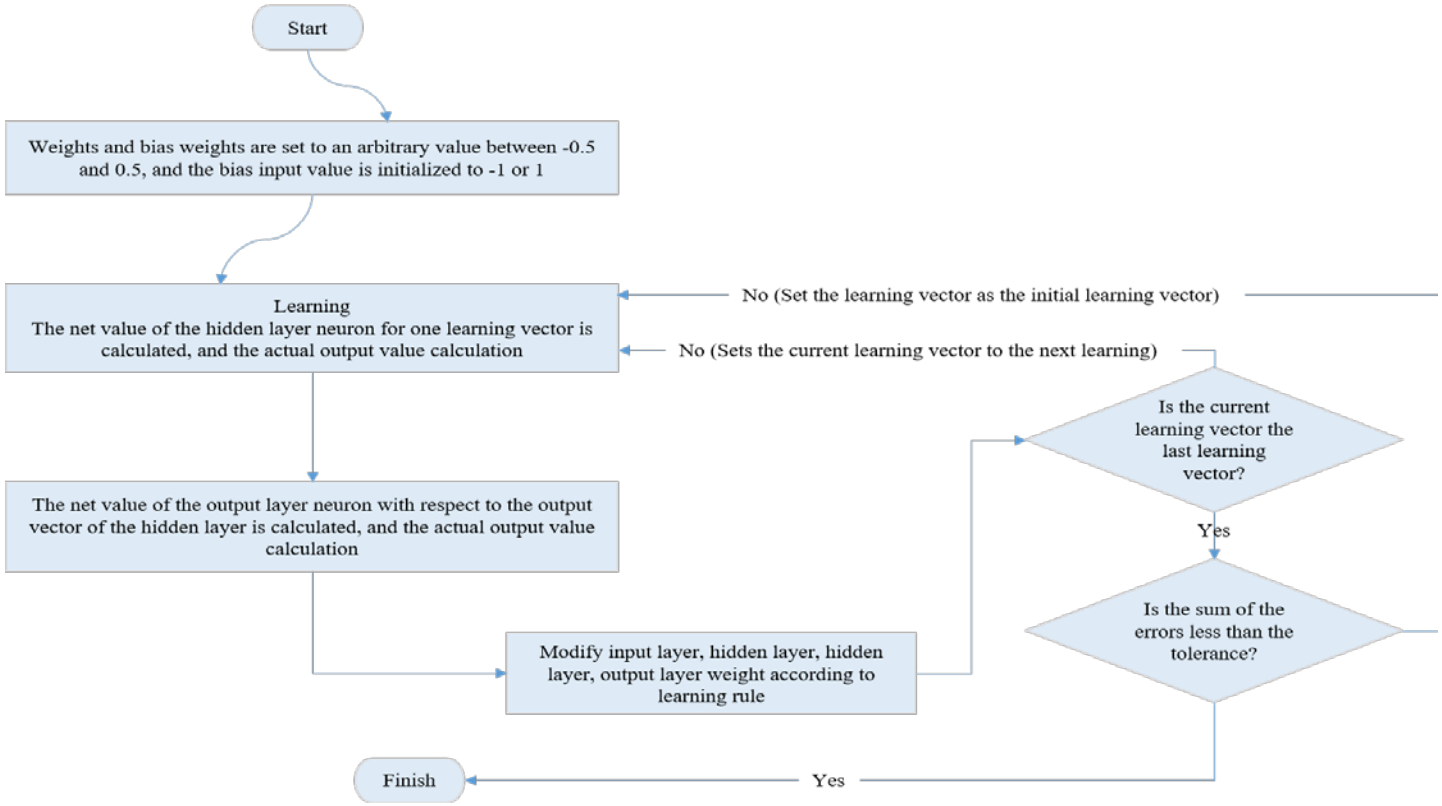


**Figure 2.3 MultiLayerPerceptron (MLP) algorithm.** The workflow of multi-layer perceptron algorithm used in this study. In a multilayer perceptron, weights of input layer-hidden layer and hidden layer-output layer must be learned. In the conventional single layer perceptron, weights are learned using the difference between the target value and the actual output value, but the target value of the hidden layer cannot be defined in the multilayer perceptron. Therefore, in the multilayer perceptron, the weight of the hidden layer is adjusted by using the error of the hidden layer propagated by the error of the output layer. This sequence of processes is called backpropagation because it is done in the opposite direction to the processing direction of the neural network.

# 제 3 장. 연구 결과

## 3.1 건강 지표 기반 질병 검색 시스템

본 연구를 위해 구축된 데이터베이스는 총 10개로 만성질환 관련 항목만을 추출하여 각 연도별로 네 개(‘2013_1’, ‘2014_1’, ‘2015’_1’, ‘2016_1’)와 머신 러닝 수행을 위해 하나로 합친 ‘DB_1’, 원시 자료의 모든 내용을 parsing 과정 없이 그대로 담은 네 개의 데이터베이스(‘2013’, ‘2014’, ‘2015’, ‘2016’), 전체를 통합시켜 구축된 ‘DB’ 가 있다. 건강 지표 기반 질병 검색 시스템의 웹 인터페이스인 LCBB-SC (Lab of Computational Biology and Bioinformatics – Symptom Checker)의 구현을 위해 사용된 데이터베이스는 총 다섯 개로, 각 연도 별 가공된 네 개의 데이터베이스는 자료 검색을 위해 사용되었고 이를 하나로 합친 데이터베이스는 머신 러닝을 통해 제시된 설문 조사 항목 증상을 선택 시 대용량의 데이터를 기반으로 관련도가 높은 증상 별 목록을 제시해주는 시스템을 구축하였다. LCBB-SC 는 현재 http://lcbb3.snu.ac.kr/LCBB-SC 를 통해 이용이 가능하며, 주 메뉴 탭은 ‘About SC’, ‘Self-Checker’, ‘Health Information’, ‘Research’ 로 구성되어 있다<Figure 3.1>. ‘About SC’을 선택하면 세개의 하위 메뉴인 ‘What is SC’, ‘Data & Analysis’, ‘Contact Us’가 나타나며, ‘What is SC’ 페이지에는 본 시스템의 간략한 소개와 사용법이 나와있다. ‘Data & Analysis’ 페이지에는 사용된 데이터의 출처와 기본 정보들, 그리고 분석 과정에 이용된 머신 러닝 알고리즘에 대한 정보를 제공하고 있다. ‘Contact Us’에는 오류나 건의 사항을 바로 관리자에게 작성해서 보낼 수 있는 페이지로 구성되었다. ‘Self-Checker’탭에는 따로 하위 메뉴가 없으며, 선택 시

바로 자가 증상 점검 프로그램 이용이 가능하도록 되어있다. 증상을 입력하기 전 먼저 간단한 사용자의 성별, 나이, 키, 몸무게 와 같은 기본 정보를 입력하고, 다음 단계에서는 혈압, 맥박, BMI, 허리둘레 정보를 입력하는 페이지가 나온다<Figure 3.2(a)>. 정보가 불확실한 경우 입력하지 않아도 다음 단계로 넘어갈 수 있으며, BMI 와 같은 수치는 신장과 체중의 정보로 자동 계산해 준다. 다음 화면에서는 현재 본인이 앓고 있는 질병이나 의사로부터 확진을 받은 질병의 여부를 선택하고, 다음 단계에서 현재 갖고 있는 증상을 선택할 수 있도록 하였다<Figure 3.2(b)>. 증상에는 원시 자료에서 제공한 '기침 발현 여부', '가래 발현 여부', '혈담 발현 여부', '흉통 발현 여부', '호흡 곤란 발현 여부', '체중 감소 발현 여부', '피로감 발현 여부', '발열 여부' '야간 발열 여부', 목록 중 선택 할 수 있고 중복 선택 또한 가능하다. 마지막 단계에서는 음주나 흡연의 여부를 선택하도록 되어있으며, 비 흡연자나 음주를 하지 않는 경우에는 바로 결과를 확인할 수 있으며, 음주자나 흡연자의 경우에는 기간, 횟수 등과 같은 세부 항목을 선택한 다음 결과를 확인할 수 있다<Figure 3.2(c)>. 'Health Information' 탭에는 '고혈압', '이상지질혈증', '뇌졸중', '심근경색', '협심증', '폐결핵', '천식', '당뇨병', 갑상선 질환', '위암', '간암', '폐암', '갑상선 암', '신부전', '간경변증'에 관한 문헌 정보를 수록하고 있다<Figure 3.3, Figure 3.4>. 마지막으로 'Research' 탭에서는 앞서 언급된 10개의 데이터베이스에 관한 정보를 확인할 수 있으며, 연구에 목적에 따라 가공된 자료를 다운로드 받을 수 있다. 또한, 본 연구에서 사용된 머신 러닝 알고리즘과 응용 가능한 Java 코드를 제공함으로서 관련 연구 인프라를 제공하고 있다. 오른쪽 상단의 검색 기능을 통해 질병 관련 정보 검색이 가능하며 추가 정보를 얻을 수 있도록 관련 링크를 함께 제시해주었다.

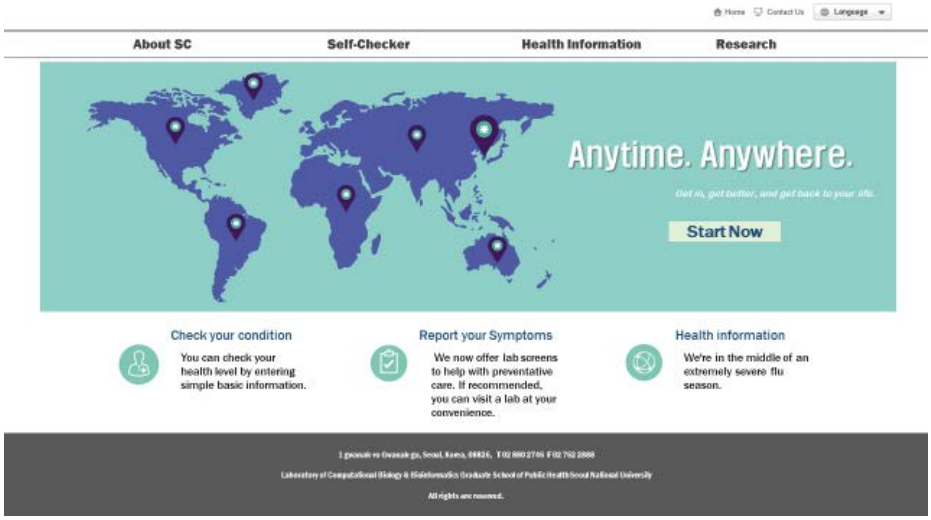


**Figure 3.1 Main page of the constructed database.** LCBB-SC (http://lcbb3.snu.ac.kr/LCBB-SC) is a web interface that provides valuable health information, tools for managing health, and support to those who seek information.

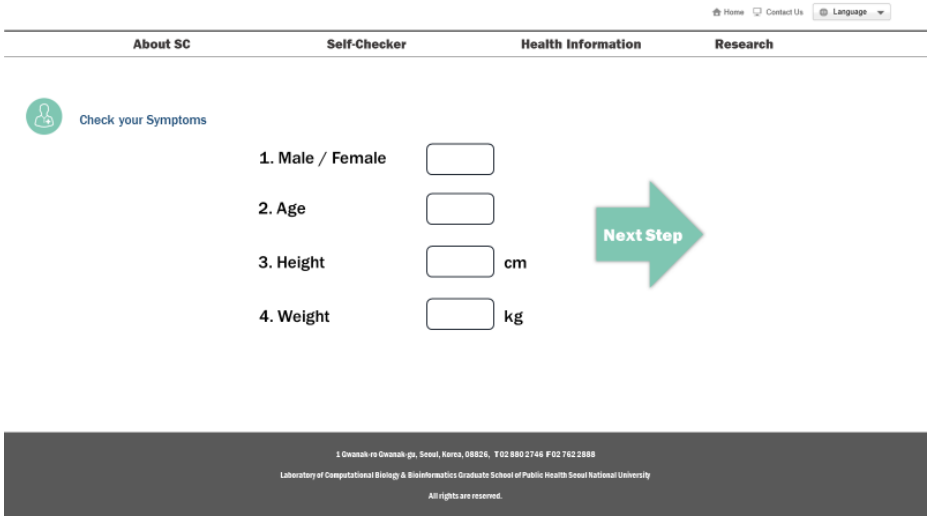


**Figure 3.3 (a) Example of self-symptom checker tool.** This is the first page of the step to start self-symptom checker. Sex, age, height and weight information might be needed to start the program.

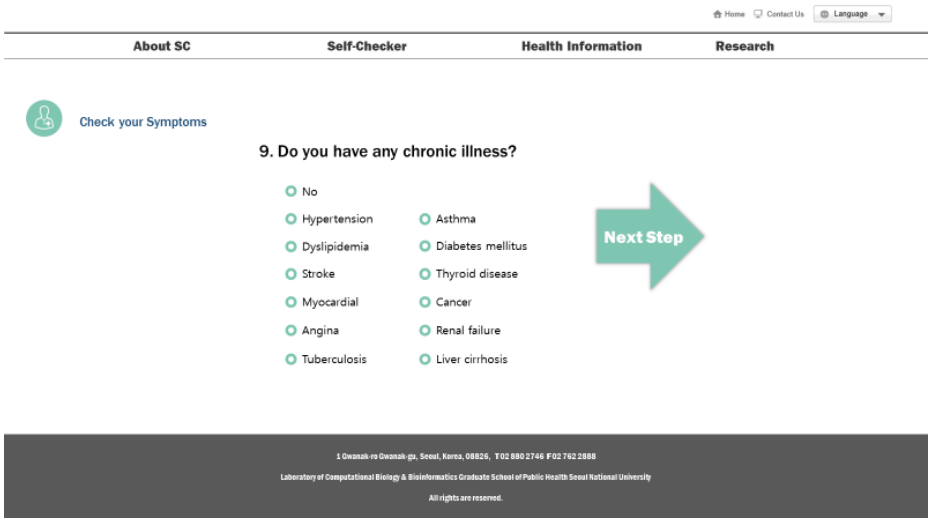


**Figure 3.2 (b) Example of self-symptom checker tool.** It is a survey page about the disease that currently have. Depending on the presence or absence of disease, different results are available and multiple selections are possible.

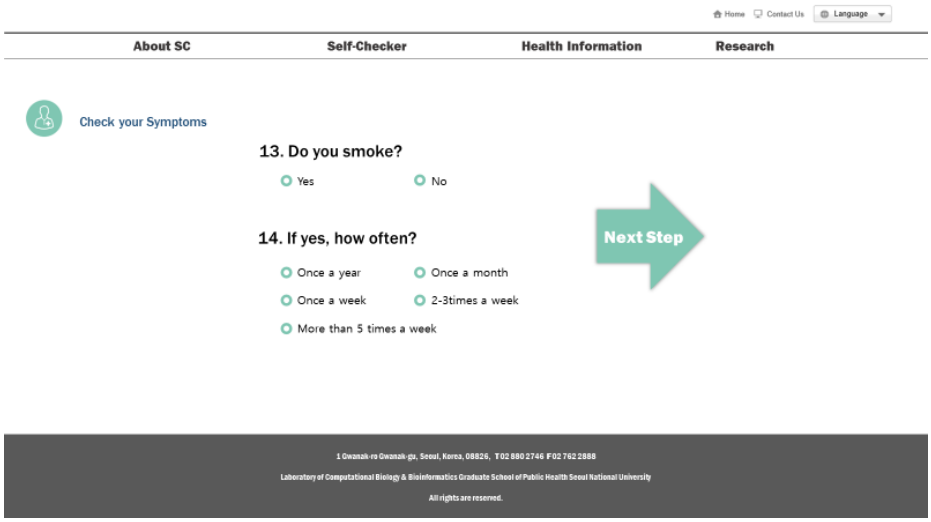


**Figure 3.2 (c) Example of self-symptom checker tool.** It is a question page about health behavior. In the case of smokers, the following questionnaire is different depending on how often they smoke, and the results of alcohol drinking also differ according to frequency.

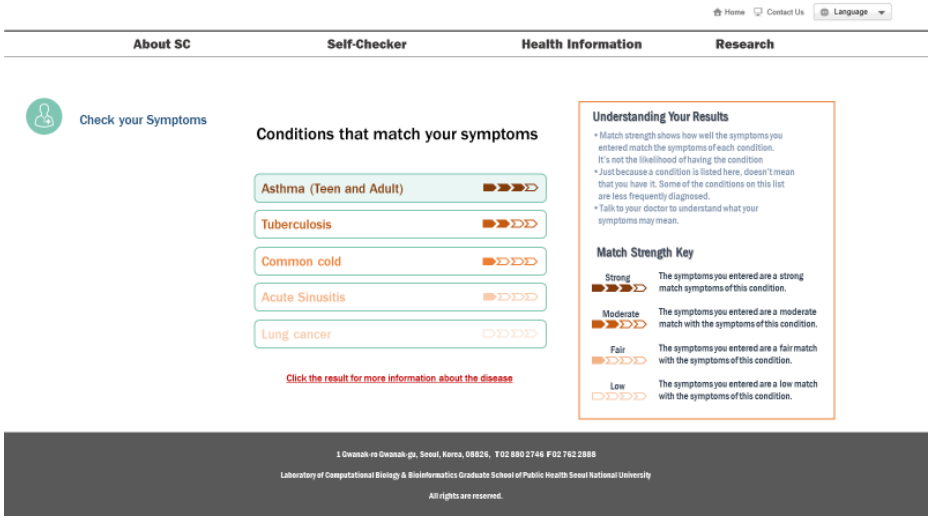


**Figure 3.3 Web interface result page.** (http://lcbb3. snu.ac.kr/LCBB-SC/result) It shows related chronic disease or conditions that match user's input information. Also, it provides detail information about the conditions by clicking the text.

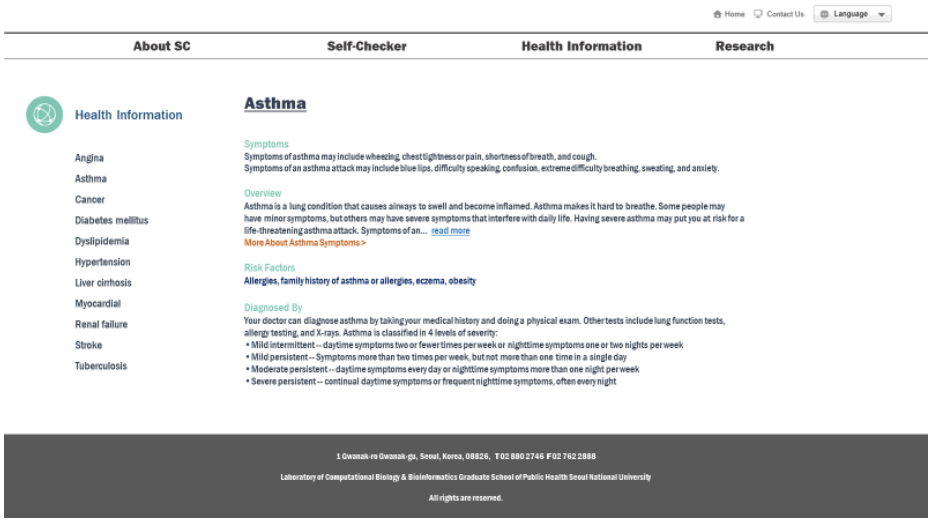


**Figure 3.4 Provide major chronic disease information.** Health information page (http://lcbb3. snu.ac.kr/LCBB-SC/info) provides valuable health information with the links.

## 3.2 머신 러닝을 활용한 분석 결과

### 3.2.1 머신 러닝을 활용한 연관성 분석

원시 자료에서 제공된 건강 지표들과 질병들 간의 상관 관계를 파악하고 예측 정확도를 비교하기 위해 Weka 3.8.2을 이용하여 분석을 시행하였다. Attributeselection 객체를 통해 evaluator 파라미터로 정보 획득을 사용하고 ranker파라미터로 정보 획득 점수를 사용하여 분석한 결과 정보성이 가장 높은 속성은 6(HE_wc), 7(HE_BMI), 16(HE_sput2), 18(HE_tb1_D), 23(HE_tb4), 24(HE_tb4_D), 30(HE_tb7_D), 34(HE_tb9_D), 38(BD1_11), 41(BD2_32), 58(DI1_dg), 71(DJ2_pr), 74(DE1_dg), 95(wt_pft) 순으로 나타났다. 좀 더 정확하고 신속한 예측을 내릴 수 있는 모델을 만들기 위해 러닝 데이터 세트에서 핵심적이지 않은, 정보성이 낮은 속성을 제거하고, 선택된 데이터를 이용하여 분류 학습 모델을 만들어 퀸란의decision tree learner를 재 구현한 J48 클래스를 사용한 결과 출력된 트리는 총 48개의 노드를 지니고 있었으며, 이들 중 22개는 종단 점으로 나타났다. ‘ActivationFunction.java’ 클래스를 통해 시그모이드 노드 48개에 따른 분석결과 net값이 임계(threshold) 값보다 큰 뉴런의 출력 값은 활성화 시키고, 그렇지 않은 뉴런의 출력 값을 비활성화 시키는 함수를 적용하였다.

### 3.2.2 인공신경망을 활용한 분석

’MLP’패키지의 ‘LinearRegression’클래스를 이용하여 각 질병별 속성 분석을 총 35회 실행한 결과 증상 및 행태, 각 질환의 의사 진단 여부와 현재 유병 여부에 따른 상관 관계를 확인하고 모델의 성능평가를 위해 10회(10-fold) 교차 평가 모델을 실행하였다<Table

3.1>. 고혈압의 경우 수축기 혈압과 이완기 혈압과의 상관 관계가 가장 크게 나타났으며, BMI (>30kg/m)와 현재 하루 평균 흡연량 과도 높은 상관 관계가 있음을 확인할 수 있었다. 만성 질환중에서는 뇌졸중, 심근경색, 협심증, 당뇨, 갑상선, 신부전과 양의 상관관계를 보였다. 이상지질혈증의 경우 역시 BMI, 음주와 높은 상관관계를 보였으며, 만성질환중에서는 뇌졸중, 당뇨, 신부전과 상관관계가 있음을 확인할 수 있었다. 뇌졸중의 경우 발열 일수 및 흡연량과 비교적 높은 상관 관계를 보였으며 고혈압, 이상지질혈증, 당뇨와 관계가 있는 것으로 나타났다. 심근경색은 흉통과 호흡곤란의 발현 여부와 높은 상관 계수를 보였으며, 만성 질환중에서는 협심증, 당뇨와 상관 관계가 있음을 확인할 수 있었다. 협심증 또한 마찬가지로 흉통의 발현 여부, 일수, 흡연과 높은 상관 관계를 보였으며, 질환중에서는 심근경색, 갑상선 질환과 당뇨의 유병 여부에 따른 높은 상관 계수를 확인할 수 있었다. 폐결핵의 경우 연속 3개월 이상 기침 여부, 호흡 곤란 발현 여부와 일수에서 양의 상관 관계를 보였다. 천식은 호흡 곤란, 간접 흡연 노출 여부와 높은 상관 계수가 나타났다. 당뇨병의 경우는 가장 많은 양의 상관 계수를 가지고 있었다. 나이, 체중, BMI, 혈압, 음주, 흡연과 같은 건강 지표를 비롯하여 대부분의 만성질환과 높은 상관 관계를 보였다. 특히 흡연, 고혈압, 이상지질혈증과 큰 상관 계수를 가지고 있는 것으로 나타났다. 갑상선은 나이와 고혈압, 협심증과 양의 상관 계수를 갖는 것으로 나타났다. 악성 신생물의 경우 음의 상관 관계가 종종 보여졌는데, 위암의 경우 발열 일수와 높은 상관 계수를 보였고 특히 성별에서 음의 상관관계를 보였는데 이는 남성에게서 많이 나타난다고 해석할 수 있다. 간암의 경우 성별, 월간 음주율과 피로감 발현 일수와 간경변증의 진단 여부에 높은 양의 상관 관계를 보인 반면, 하루평균 수면시간과 2주 이상 연속

우울 감 여부와 음의 상관 관계를 보였다. 폐암의 경우 최근 2주간 몸이 불편했던 경험의 유무, 음주, 직장내 간접흡연 유무, 갑상선 암, 평소 스트레스 인지 정도, 평균 수면시간과 높은 상관 관계가 있는 것으로 나타났다. 갑상선 암의 경우 나이, 음주, 심근경색, 고혈압과 높은 상관 관계가 나타났으며, 신부전의 경우 피로감 발현일 수, 심근경색 및 협심증, 고혈압, 이상지질혈증과 상관 관계가 높게 나타났으며, 간경변증의 경우 간암과의 양의 상관 관계만 나타났다.

**Table 3.1 Multiple layer perceptron result for chronic disease (10-folds)**

| Code | Data (No. of data) | Code | Data (No. of data) |
|---|---|---|---|
| DI1_dg | 6.3455 * BS3_2 +<br>2.2548 * DN1_dg +<br>2.2643 * DN1_pr +<br>11.2264 * HE_sbp1 +<br>10.9765 * HE_dbp1 +<br>10.7784 * HE_BMI +<br>8.4531 * DI4_pr +<br>5.5124 * DE1_dg<br>4.4894 * DE1_pr | DI6_dg | 20.9892 * DI4_dg +<br>8.8985 * HE_tb4 +<br>17.7637 * DI5_dg +<br>9.5541 * HE_tb4_d +<br>4.3039 * BS3_1 +<br>6.4484 * BS3_2 +<br>6.8841 * DE2_dg<br>4.1142 * DE1_dg |
| DI2_dg | 2.2131 * DI4_pr +<br>-2.2208 * DC6_pr+<br>7.6924 * HE_BMI +<br>6.9944 * dr_month +<br>4.2927 * DI3_pr +<br>6.2867 * DE1_dg<br>5.4841 * DE1_pr | DJ2_dg | -2.2545 * DI4_pr<br>2.2961 * DI5_pr +<br>9.1154 * HE_cough1<br>9.3122 * HE_tb5 +<br>7.4487 * HE_tb5_d |
| DI3_dg | -7.3683 * HE_tb8<br>4.3205 * HE_tb9 +<br>7.3455 * BS3_2 +<br>7.2667 * DI1_dg +<br>5.9564 * DI2_dg +<br>4.5675 * DE1_dg | DJ4_dg | 6.4167 * HE_tb5 +<br>4.8871 * BS8_2 +<br>3.1451 * BS9_2 +<br>4.5771 * BS13 + |
| DI4_dg | 0.8016 * DI5_dg +<br>0.9171 * DI6_dg + | DE1_dg | 6.4286 * age +<br>8.2416 * HE_wt +<br>9.7784 * HE_BMI +<br>5.2264 * HE_sbp1 +<br>3.3714 * HE_dbp1 +<br>2.4197 * BD13 +<br>4.5417 * dr_month +<br>8.8743 * BS3_2 +<br>2.4875 * BS6_2 +<br>8.4495 * DI1_dg +<br>6.7314 * DI2_dg +<br>3.7112 * DI3_dg +<br>2.4495 * DI5_dg +<br>1.9411 * DI3_dg + |
| DI5_dg | 22.4495 * DI4_dg +<br>16.2391 * DI6_dg +<br>10.8972 * HE_tb4 +<br>7.9632 * HE_tb4_d +<br>9.6544 * HE_tb5 +<br>6.3554 * HE_tb5_d +<br>6.4541 * DE1_dg | DE2_dg | 4.5495 * HE_tb6 +<br>7.9614 * age +<br>6.3412 * DI1_dg +<br>2.7319 * DI6_dg + |

**Table 3.1 Multiple layer perceptron result for chronic disease (10-folds) (cont.)**

| Code | Data (No. of data) | Code | Data (No. of data) |
|---|---|---|---|
| DC1_dg | 6.4167 * HE_tb8 +<br>-4.9784 * sex<br>3.0547 * BD1_11 + | DC7_dg | 8.3645 * age +<br>2.9465 * dr_month +<br>3.4998 * DI5_dg +<br>2.0107 * DI1_dg_2 +<br>2.0013 * DE1_dg + |
| DC2_dg | 4.7984 * dr_month +<br>3.1194 * HE_tb7_d +<br>2.1974 * DK4_dg +<br>-6.1173 * sex+<br>-2.4855 * BP8<br>-1.5487 * BP8 | DN1_dg | 2.9545 * HE_tb7 +<br>1.9451 * DI5_dg +<br>2.5498 * DI6_dg +<br>2.0641 * DI1_dg +<br>1.7498 * DI2_dg + |
| DC6_dg | 2.7984 * D_2_1 +<br>6.7064 * BS8_2 +<br>3.9322 * dr_month +<br>2.0494 * DC7_dg + | DK4_dg | 3.3316 * DC_2dg + |

Using the 'LinearRegression' class, 35 regression analyzes for each disease. The correlation between the diagnosis of each disease and the current prevalence was confirmed, and 10-folds crossover evaluation was performed to evaluate the performance of the model. Each code name uses the variable name used in the source data, and the explanation is given in Table 2.1.

## 3.3 상관분석과 거리행렬

구축된 데이터베이스 중 4차년도 통합 데이터베이스를 로드하고 SAS를 활용하여 다중 상관 분석을 실행하였다. 설명 변수의 관측값을 이용하여 반응 변수의 관측값을 예측하는 회귀분석과는 달리 상관분석은 변수들 사이의 연관성의 크기를 찾는 통계분석 방법이며, 다중상관계수는 다중결정계수의 제곱근으로 나타낼 수 있다. 분석 결과 중 통계적으로 유의한 차이를 갖는 (p-value<0.0001) 값만을 추출하였으며, 이를 바탕으로 distance matrix를 작성하였다. 신경망 알고리즘을 활용한 분석 결과와 비교하였을 때 상당 수가 일치하였으며, 상관계수가 0에 가까운 값은 신경망 알고리즘 분석 결과에서는 나타나지 않았음을 확인 할 수 있었다. 상관분석의 결과 값을 이용하여 머신 러닝 분석과의 예측도를 비교해본 결과 알고리즘을 활용한 분석의 경우 예측률이 고혈압 4.7%, 뇌졸중 9.8%, 심근경색 3.2%, 당뇨 11.3% 높게 나온 것을 확인할 수 있었다. 또한 협심증의 경우 2.1% 정도 높은 예측률을 보였지만 각종 악성 신생물의 경우는 거의 비슷한 예측률이 나오는 것을 확인 할 수 있었다. 암 환자의 경우 데이터의 양이 많지 않기 때문에 기존의 분석 방법과의 차이가 확연히 드러나지 않는 반면, 당뇨나 고혈압의 경우 확진 판정을 받은 데이터가 다른 질병에 비해 비교적 많이 있었기 때문에 좀 더 높은 예측률이 나타난것으로 유추해 볼 수 있다. 또한 상관분석의 경우 파라미터와 가중치 값이 조정되기 전 결과로 해석할 수 있으며, 피드포워드 기법을 적용하여 모델 학습을 실행한 결과 가중치 값이 조절되고 노이즈가 섞인 데이터 분류를 통해 좀 더 정확하고 비중있는 가중치 값만을 활용하여 분석을 실행한 결과라고 유추해 볼 수 있다.

# 제 4 장. 고 찰

## 4.1 국내 연구 현황

과학 기술과 의술의 발달로 기대수명이 늘어나면서 전 세계적으로 만성 질환의 유병률과 발병률 또한 매년 꾸준히 증가하고 있다. 특히 면역력이 취약한 노인 계층의 경우 단일 질환 유병률의 감소와 동시에 두개 이상의 만성질환을 보유한 환자의 수가 급속도로 증가하고 있다. 과거에는 전염성 질병이나 미생물에 의한 감염성 질환이 국제 공중 보건의 주요 안건이었지만, 21 세기에 들어서면서 많은 연구가 진행되어 왔으며 그 성과로 예방 백신과 치료제가 개발되었다. 하지만 만성 질환의 경우 치료기간이 길고 과거로부터 오랜 연구에도 불구하고 아직까지 직접적인 원인이 밝혀진 것이 거의 없으며 예방이나 완치가 어렵기 때문에 특히 노인들에게서 발병하는 질병의 경우 주요 사망 원인으로 보고되고 있어 고령화 시대의 공중보건학적으로 특히 중요한 문제이다. 또한 사회적으로 취약 계층이나 빈곤층의 경우 환경이나 경제적인 문제로 제때 치료를 받지 못해 단순 질병이 악화되어 만성으로 진행되거나 각종 합병증을 겪고 있어 해결이 시급한 실정이다. 뿐만 아니라 대부분의 만성질환의 경우 초기 증상이 거의 나타나지 않거나 미미하여 진료를 받지 않고 가벼이 넘어가기 때문에 병원을 찾는 질환자의 대다수는 이미 질병이 치료 수준을 넘어 상당히 진행된 상태이거나 다른 신체 부위로 전이가 된 상태이다.

본 연구를 수행하기에 앞서 많은 만성 질환 관련 연구를 조사해본 결과 대부분 단일 질병에 관한 연구로 주로 질병 요인을

찾거나 질환 간 상관 관계를 밝혀내는 연구가 압도적으로 많았다. 최근 들어 머신 러닝 기술이 주목받기 시작하면서 이를 활용한 만성 질환 연구가 진행중이지만 아직 시작 단계에 불과한 것을 확인할 수 있었다. 특히 국내의 경우 조사된 자료는 많지만 아직까지 관련 연구자들이 연구에 사용할 수 있도록 체계적인 분류를 통해 비교적 정리되어있는 정보를 제공하고 있는 플랫폼은 부족한 실정이다. 연구자가 원하는 데이터를 수집하기에 검색이 어렵거나 중복된 데이터가 다수 존재하며, 연구자가 원하는 정보만을 얻기 위해 수집한 데이터를 다시 가공하여 정리해야하는 어려움이 따른다. 또한, 해외의 사례와 비교하였을 때 신뢰할 만한 연구를 바탕으로 관련 종사자가 아닌 일반인들에게 만성 질환에 대한 사전 지식을 제공하며 자가 점검을 통해 질환 발병률을 낮추는데 기여할 수 있는 국내 보건 시스템이 부족한 상황이다. 따라서 본 연구에서는 4차년도 설문 조사 자료를 바탕으로 만성 질환과 관련 요인들을 선정하여 데이터베이스를 구축하여 관련 질환 연구자들에게 유용한 데이터를 제공하고자 한다. 추후 유전자 정보 및 식습관 데이터를 추가하여 다양한 연구의 기반이 되는 인프라를 구축하고자 한다. 또한, 정보 탐색과 새로운 연구 결과를 바탕으로 지속적인 업데이트 과정을 통해 활용 가치가 높은 데이터를 제공하고자 하였으며, 주기적인 유지 및 보수 작업을 실행하고자 한다. 아울러 분석 결과를 통해 얻어진 건강 지표를 바탕으로 질병 예측률을 계산하고 이를 온라인 자가 점검 시스템에 접목시켜 개방할 예정이다. 이를 통해 일반인들에게 고 위험성 만성 질환 관련 위험 요인들과 질병에 대한 사전 정보 및 관련 최신 연구를 소개함으로써 해당 질병의 발병률과 유병률을 낮추어 국민 보건에 기여하고자 한다.

## 4.2 질병 및 건강 지표의 연관성

먼저 구축한 LCBB-SC의 데이터를 활용한 머신 러닝 분석 결과를 살펴보면, 고혈압의 경우 측정 혈압을 제외한 항목 중 BMI가 미치는 영향은 10.7784로 가장 높았으며, 하루 평균 흡연량 또한 6.3455로 높은 영향을 주는 것으로 나타났다. 하버드대학 연구팀의 연구 결과에 따르면 비만뿐 아니라 BMI 수치가 높을수록 고혈압 발병 위험도가 높아진다고 밝혔다. 이들은 고혈압이 없는 성인 남성 13,563명을 대상으로 장기간 추적 연구를 진행한 결과 BMI 5분위 최하위군(22.4㎏/㎡ 미만)과 비교한 고혈압 발병의 상대적 위험은 BMI가 22.4～23.6㎏/㎡에서는 1.20, 23.7～24.7㎏/㎡에서는 1.31, 24.8～26.4㎏/㎡에서는 1.56, 26.4㎏/㎡을 초과한 경우에는 1.85로 모두 유의하게 높았다(P＜0.0001). 이러한 관련성은 당뇨병이나 높은 콜레스테롤수치, 등록 당시 혈압을 조정해도 거의 바뀌지 않았으며 BMI가 정상 및 경미한 과체중 범위 내라도 BMI의 상승은 고혈압 위험과 밀접하게 관련한다는 것을 나타낸다(Gelber et al., 2007). 흡연자의 경우 담배에 있는 니코틴이 뇌에서 감지되면 아드레날린을 분비하게 되고 이로 인해 혈관이 수축되며 심박수를 빠르게 촉진시키게 된다. 또한 담배 속에 있는 각종 화학 물질로 인해 직접적인 혈관 손상도 초래하기 때문에 고혈압을 악화 시키는 주요 요인으로 밝혀져 있다. 실제 담배 2개피를 연달아 피운 결과 혈압이 10mmHg상승하는 것을 확인할 수 있었으며 이러한 상태가 30분 이상 지속되었다(Polosa et al., 2016). 고혈압과 높은 연관성이 있는 질환으로는 심근경색 또는 협심증이 8.4531로 가장 높았으며 당뇨(5.5124), 신부전(2.2643), 뇌졸중(2.2548) 순으로 나타났다. 실제 임상 연구 결과에 의하면 고혈압의 합병증으로 심장 질환이 가장 많았으며 전체 고혈압 환자의 39%가 당뇨를 가지고 있는 것으로

밝혀졌다. 과거에는 신부전과 고혈압의 직접적인 연관성 보다는 당뇨병의 합병증으로 밝혀진 연구 사례가 더 많았지만(Bakris et al., 2000; Klausen et al., 2004), 최근에는 고혈압에 의한 신부전 발병 가능성이 제시되고 있다(Hall et al., 2015; Xie et al., 2016). 이상지질혈증의 경우 역시 BMI가 7.6924로 가장 큰 영향을 미치며 음주(6.9944)와 높은 관련성이 있다는 결과가 나왔다. 특히 최근 들어 중국에서 많은 연구가 이루어지고 있음을 확인할 수 있었는데(Yu and Sun, 2015; Li et la., 2017; Xin et al., 2018), 상대적으로 기름진 음식을 많이 섭취하며 특히, 활동량이 적은 중국의 여성에게서 흔히 나타나는 질병으로 BMI 수치와 밀접한 관계가 있는 것을 확인할 수 있었다(Miura et al., 2017). 과거의 한 연구에서는 하루에 30 g의 에탄올을 섭취하는 경우에 평균 고밀도지질단백(High-Density Lipoprotein, HDL)의 혈중농도가 3.99 mg/dL가 증가하고, 아포단백 A-1 (Apoprotein A-1)의 농도가 8.82 mg/dL 상승한 다고 보고하였다(Rimm et al., 1999). 또한 하루 30 g을 초과해서 에탄올을 섭취하면 중성지방의 농도를 상승시킬 수 있으며, 하루 60 g을 섭취하는 경우 에탄올 1 g당 중성지방을 0.19 mg/dL 상승시킬 수 있다는 보고도 있었다(Oh, 2009). 최근 국내의 연구 결과에 의하면 폭음, 잦은 술자리 등과 같은 고위험 음주를 하는 남녀 성인의 경우, 음주 습관이 나쁘지 않은 성인보다 저밀도 지질 단백질-콜레스테롤(low density lipoprotein cholesterol, LDL-C) 위험성이 남성 1.5배, 여성 1.6배 증가하였으며 남성의 경우 고콜레스테롤혈증과 고중성지방혈증이 1.2배, 2.0배 가량 높아지는 것을 확인할 수 있었다(Kwon et al., 2016). 이미 흡연량과 뇌졸중에 대한 많은 연구가 이루어져 왔기 때문에 본 연구 분석 결과에서의 높은 수치(7.3455)는 비교적 타당성이 있는 결과로 볼 수 있다. 하지만 발열 일수와는 음의 관계(-7.3686)를 보였는데, 이는 발열 일수가 적거나 없는 경우

뇌졸중의 발병률이 증가한다고 해석할 수 있지만 이를 뒷받침해 줄 근거나 연구 사례는 찾아볼 수 가 없었다. 따라서 이에 관한 추가적인 문헌 조사나 해당 부분의 알고리즘을 수정해야할 필요가 있다고 사료된다. 뇌졸중 관련 만성질환으로는 고혈압(7.2667), 이상지질혈증(5.9564), 당뇨(4.5675) 순으로 나타나는데, 최근 캐나다에서 약 2만7000명을 뇌졸중 환자 13,447명과 대조군13,472명으로 나누어 비교하여 뇌졸중 발생 시 기여하는 위험 정도를 분석한 결과 뇌졸중의 가장 큰 위험 요소는 고혈압으로 뇌졸중 발생에 47.9% 기여하는 것으로 나타났다(Govindarajan et al., 2017). 뇌는 우리 몸에서 가장 혈류가 많이 가는 장기이므로 혈압의 영향에 특히 민감하기 때문이라고 해석할 수 있으며, 본 연구의 분석결과에 뒷받침이되는 근거로 볼 수 있다. 심근경색과 협심증은 22.4495와 20.9892로 상호 관계성이 상당히 높은 질환이며, 이 두 질환에 관련된 위험 요소 또한 호흡곤란(10.8972 / 8.8985), 흉통(7.9632 / 9.5541), 흡연(17.7637)으로 유사한 것을 확인할 수 있었다. 공통 만성 질환으로는 당뇨가 각각 6.4541과 4.1142로 다른 질환에 비해 높은 연관성을 나타내는 것을 확인할 수 있었다. 참고 문헌 조사 결과 당뇨와 고혈압에 의한 혈관 손상으로 심근경색과 협심증이 발생할 수 있으며, 실제 국내 당뇨, 고혈압 환자를 대상으로 조사한 결과 이들은 정상인들에 비해 심근경색 발병률이 6배나 높은 것으로 나타났다(Won et al., 2015). 폐결핵의 경우 기침 발열 일수(9.1154)와 호흡 곤란 발현 여부(9.3122)가 가장 큰 영향을 미치는 것으로 나타났다. 하지만 폐에 직접적인 영향을 주는 요인인 흡연의 여부는 큰 영향을 못 미치는 것으로 나타났으며, 폐암과는 거의 연관성이 없다는 결과가 나왔다. 문헌 조사 결과, 아직까지 흡연이 폐결핵 질환에 직접적인 영향을 준다는 과학적인 근거를 제시하는 연구는 없었으며, 몇몇 보고에 의하면 폐결핵이 폐암을

유발하는 경우도 있었지만(Radzikowska et al., 2002; Cicenas and Vencevicius, 2007), 폐결핵 환자가 폐암의 발병 위험성이 높다고 판단하는 것에는 아직 무리가 있는 것으로 보여진다. 당뇨는 이미 잘 알려진 대표 만성질환으로 관리가 잘 이루어지지 않을 시 단일 질병 상태로 남아 있기 보다는 하나 이상의 합병증을 유발할 수 있는 질병이다. 당뇨가 오래 지속될 경우 혈관이 좁아지거나 막히게 되면서 다른 질환을 초래할 수 있으며(Schror, 1997), 본 연구 결과에서도 당뇨의 경우 가장 많은 요인들과의 연관성이 있는 것을 확인할 수 있다. 분석 결과에 의하면 특히 BMI(9.7784), 몸무게(8.2416), 나이(6.4286)에 많은 영향을 받으며, 월간 음주량(4.5417)과 흡연량(2.4875) 또한 주요 요인으로 작용한다. 질병관리본부의 2015년 4월 발표한 보고서에 흡연과 당뇨병 사이의 연관성을 비교 분석한 결과에 따르면 비 흡연 군에 비해 하루 한 갑(20개비) 이상 담배를 피우는 흡연 군에서 당뇨병 위험이 1.54배 높은 것으로 나타났다. 또한 음주 시 체내에 흡수된 알코올은 췌장에서의 인슐린 분비를 막고, 동시에 간 기능 장애를 일으켜 혈당을 높인다는 연구 결과가 보고되었다(Wakabayashi et al., 2002). 갑상선 질환의 경우 나이가 7.9614로 가장 큰 영향을 미치는 것으로 나타났으며 체중 감소 여부 또한 4.5495로 양의 상관관계가 있음을 알 수 있다. 관련 만성 질환으로는 고혈압이 6.3412로 가장 높았으며 협심증은 2.7319로 비교적 낮았지만 다른 만성질환에 비해 상관관계가 큰 것으로 나타났다. 갑상선과 고혈압이나 심장질환에 관련성은 많이 언급이 되지 않았기 때문에 연관성에 관한 문헌을 찾는 것은 쉽지 않았기 때문에 알고리즘이나 설문 조사 자료에 의한 오류일 것이라 추측되었다. 하지만 최근 유타 대학의 한 연구팀의 연구 결과 발표에 따르면 40세 이전에 갑상선 질환 진단을 받은 환자 대부분은 고 용량의 약물을 처방 받게 되면서

이를 장기적으로 시행하게 될 경우 고혈압, 심혈관계 질환 등에 문제가 생기는 등의 부작용이 발생하는 것으로 추정된다고 밝혔다(Blackburn et al., 2017). 위암의 경우 성별에서 -4.9784로 음의 상관 계수가 나타났는데 이는 남성의 성별은 1, 여성은 2로 설정되어있는 것을 바탕으로 볼 때 남성에게서 주로 많이 나타나는 질병이라고 해석할 수 있다. 본 연구 결과에서는 언급되지 않았지만 음주 빈도와 성별 또한 음의 상관 관계를 보였는데, 남성의 경우 여성에 비해 평균 음주량이 높으며 이로 인하여 위암에 성별이 간접적인 영향을 미치는 것으로 결론지을 수 있다. 실제 음주와 위암과의 상관성은 과거 수많은 연구와 최근 메타 분석 연구에 의해 밝혀졌기 때문에(Chow et al., 1999; Boffetta and Hashibe, 2006; Ma et al., 2017), 본 연구 결과에서 나타난 성별 또한 위암의 중요한 건강 지표로 사용될 수 있다. 흥미로운 점은 간암의 경우 음주량, 피로감과 같은 직접적인 영향을 주는 것으로 이미 잘 알려진 결과 외에, 2주 이상 연속 우울 감 여부와 음의 상관 관계를 보인 것은 음주 빈도가 높을수록 우울 감은 낮다는 것을 의미한다. 하지만 기존 연구결과에 따르면 과음은 간을 비롯한 몸의 여러 기관, 특히 뇌에 악영향을 주게 되는데, 이는 체내로 흡수된 알코올이 분해되면서 만들어지는 아세트알데히드가 뇌 조직의 손상을 유발하는 것으로 나타났다(Hasin et al., 2005). 과음 후 알코올이 분해되면서 그만큼 아세트알데히드가 뇌 조직 등 몸에 쌓이게 되고 이로 인해 쉽게 우울 감을 나타나는 것으로 우울한 감정 역시 뇌 조직이 손상 받는 급성 증상으로 볼 수 있다(Brook et al., 2002). 폐암의 경우 직장내 간접흡연 유무가 미치는 영향이 6.7064로 가장 높았다. 미국 EPA (Environmental Protection Agency)는 간접 흡연에 대한 기존의 연구를 종합적으로 분석하여 간접 흡연이 폐암의 위험인자라고 발표하였으며, 간접 흡연을 직접 흡연과는 별도로

그룹 A 발암 원으로 규정하였다(https://www.epa.gov/indoor-air-quality-iaq/ respire tory-health-effects-passive-smoking-lung-cancer-and-other-disorders). 과거 한 연구에 따르면, 600명 이상의 폐암 환자와 1,200명 이상의 대조군을 이용한 연구에서 직장 내 간접 흡연으로 인하여 폐암의 위험이 1.4 배 정도 증가한다고 보고하였다. 이 연구에 의하면 직장에서 간접 흡연에 노출된 기간도 상관 관계가 있어, 30년 이상 노출되었을 때는 상대 위험도가 2.08 (95% CI: 1.35~3.20)로 높게 나타났다(Fontham et al., 1994) 간접 흡연으로 인한 발암물질에 관한 연구나 동물 연구, 여러 역학적 연구들은 간접 흡연이 폐암의 위험 요인임을 잘 보여 주고 있으며(Besaratinia and Pfeifer, 2008; Thun et al., 2008; Samet et al., 2009), 특히 흡연율이 높은 국내에서 비흡연자의 간접 흡연에 의한 피해는 서구 국가들보다 더 심각할 수 있다. 갑상선 암은 나이가 미치는 영향이 8.3645로 가장 높게 나타났다. 아직까지 고 연령 자체가 갑상선 암의 발생을 증가시키는지는 명확하게 밝혀지지 않았지만, 외국의 경우 갑상선 암의 발생률은 나이가 많아지면서 점차 증가한다는 연구 결과가 꾸준히 보고되고 있다(Botrugno et al., 2011; Toniato et al., 2011). 또한 국내 통계에 따르면 남녀 모두 나이가 증가할수록 갑상선 암의 비율이 증가 하며, 특히 남성에서 갑상선 암의 비율은 65세 이후 비약 적으로 높아져, 75세에서 60%에 달하는 것으로 나타났다(Kim et al., 2012). 신부전의 경우 피로감 발현일 수가 2.9545로 가장 큰 영향을 미치는 것으로 나타났으며 고혈압과 당뇨 또한 2.0641, 2.0013으로 높게 나타났다. 신부전의 경우 다른 만성질환에 비해 특히 초기에는 거의 증상이 없다가 신장 기능의 저하 정도가 심해지면 야뇨증, 수면장애, 피로감, 소화장애, 구역, 구토, 소변량 감소, 부종, 기억력 감퇴, 호흡곤란, 심 부정맥과 같은 증상이 나타난다(Chronic Kidney Disease Prognosis Consortium, 2010). 현재까지 밝혀진 신부전의 원인은 전

세계적으로 당뇨병성 신증이 가장 흔하며 고혈압성 신경화증과 만성 사구체 신염이 2,3위로 밝혀져 있다 (Nahas and Bello, 2005; Levey and Coresh, 2012).

기존 만성 질환과 관련된 대용량의 데이터를 바탕으로 본 연구에서 수행한 분석 결과는 건강 지표를 이용하여 관련 만성 질환의 예측을 위한 적합한 알고리즘 선택의 근거로써 제시될 수 있으며, 이를 수치화 하여 가중치로 사용된 신경망의 타당성 있는 예측 결과를 도출할 수 있다고 사료된다. 또한 질병 위험 요인 분석 뿐 아니라 만성 질환 간의 수치화된 연관성 결과를 제시하였다는 점이 본 연구의 차별화된 전략이라고 할 수 있다. 향후 국민 건강 영양조사에서 업데이트되는 데이터와 본 연구에서는 포함되지 않았던 식습관, 유전적 데이터 등을 추가하여 만성 질환 관련 연구에 좀 더 다양하고 신뢰할 만한 인프라를 제공하고자 한다. 앞서 언급한 갑상선 질환과 고혈압, 심혈관 질환과 같이 연구에서는 밝혀지지 않은 새로운 질환 간의 상관 관계를 유추하거나 예측하는데 유용하게 사용될 수 있다. 또한 분석 과정에 쓰인 머신 러닝 알고리즘을 문헌 조사 결과나 전문 서적을 통해 오류로 판단되는 분석 결과를 보완하여 궁극적으로 신뢰할만한 독자적인 국내 대규모 만성 질환 데이터베이스 구축을 통해 보건학적 연구의 기반을 마련하고 확충하는데 이바지할 것으로 기대한다.

## 4.3 연구의 한계점

본 연구는 일반인을 대상으로 한 설문 조사 데이터를 바탕으로 분석을 실시하였다. 구조화된 설문지는 답변을 명백하고 용이하게 계량화할 수 있다는 장점이 있지만 제시된 선택 항목이 충분히 포괄적이지 못하며 모든 응답이 문항에 쉽게 대응되지 못할 수 있다는 단점이 있다. 응답자는 자신의 견해를 완전히 대표하지 못하는 경우가 발생하며, 무응답은 유효 표본 수를 감소시키므로 자료의 질에 영향을 주어 조사 추정치의 정확도를 떨어뜨린다. 따라서 실제 임상 데이터에 비해 정확도나 신뢰도가 떨어진다는 단점이 있다. 또한 분석 과정에서 사용된 머신 러닝 알고리즘은 데이터 전처리 과정이 매우 중요하며, 데이터 집합 속에 누락된 값이 있을 경우 제대로 작동하지 않을 수 있다. 따라서 본격적인 분석을 시행하기에 앞서 정확하지 않거나 불충분한, 관련성이 떨어지는 데이터를 제거해야 하며, 수용 가능한 데이터 타입인지 확인하기 위한 데이터 스트링에 대한 검증 등의 작업을 수행해야한다. 이러한 과정을 세밀하고 정확하게 시행하지 않고 분석을 실행하였을 경우 다음과 같은 문제점들이 발생할 수 있다. 첫째, 노이즈가 섞여있는 학습용 데이터 셋을 분류해 잘못된 최적화를 하는 경우이다. 특히 본 연구에서 사용된 데이터는 일반인들을 대상으로 한 설문조사 자료로 특정한 특정 정보를 제외하고는 본인의 주관적인 생각을 바탕으로 하는 응답이 대다수이므로 노이즈가 많이 섞여있어 완벽한 패턴들로 제대로 분류하는 것이 불가능하다. 머신 러닝 파라미터들을 조정하지 않고 그대로 학습 시키면 모델은 노이즈 데이터를 패턴으로 분류하게 되며, 이것을 검증용 데이터 셋에 적용할 경우 그 노이즈는 존재하지 않으므로 검증에서의 예측력은 떨어질 수밖에 없다.

이러한 경우 규칙성을 찾아내는 것이 어렵기 때문에 데이터의 질보다는 양에 의존하여 연구를 시행해야 하거나, 검증 횟수를 늘려 해결할 수 있다. 둘째, 학습용 데이터 셋에서 만의 특징적인 데이터를 분류해 잘못된 최적화를 하게 되는 경우이다. 본 연구에서는 2016년도 1차, 2차, 3차를 학습용으로, 2017년도 1차를 검증용 데이터 셋으로 나누어 분석을 실행하였다. 2016년도에 최적화된 모델은 학습용 데이터에서는 좋은 결과를 얻을 수 있어도, 검증용 데이터셋에서는 좋은 결과를 얻을 수 없으며 실제 적용에도 유용하지 않을 수 있다. 따라서 교차 검증을 통해 이러한 과 적합 문제를 해결하며 시스템을 보완해 나가는 작업이 필요하다. 은닉층의 수를 증가시킴으로써 신경망은 전체적으로 더 다양한 패턴으로 표현될 수 있으며 세분화 작업을 통해 정확도를 높일 수 있을 것으로 기대하였다. 하지만 실제 결과는 큰 차이가 없었으며, 오히려 층의 수가 많은 신경망의 예측 정확도가 적은 층의 신경망보다 떨어지는 경우도 있다는 것이 밝혀졌다(Sugomori, 2016). 이러한 문제가 발생하는 이유는 피드포워드 망의 역 전파 알고리즘이 가지고있는 특징인 때문인데, 출력층에서의 오차를 입력층으로 역전파함으로써 망의 가중치를 각 층에서 순서대로 조정하고 망의 전체 가중치를 최적화하는 과정에서 신경망의 층 수가 적다면 출력 층의 오차 역 전파는 각 층의 가중치를 원활히 조정하는 데 기여할 수 있지만 층의 수가 증가하면 역 전파가 각 층을 통과하면서 오차가 점차 희석돼 사라져 버리게 되어 결국 전체 망의 가중치를 조정하지 못하게 된다(Sugomori, 2016). 또한 층의 수가 많아질수록 오차 또한 많아지게 되며 각 링크의 가중치들이 분산되어 반영되는 몫이 작아지게 되어 가중치 조절이 어려워진다는 단점이 생긴다. 따라서 망 전체의 가중치를 최적화할 수 없으며, 결국 신경망 전체의 예측률이 떨어지게 된다.

## 4.4 활용 방안 및 기대 성과

본 연구에서는 국민건강영양조사에서 제공하는 설문 데이터를 활용하여 주요 만성 질환에 특화된 데이터베이스를 생성하고, 이를 바탕으로 분석 과정을 통해 건강 지표를 찾고 머신 러닝 기법을 기반으로 한 질병 예측 시스템을 구축하였다. 이를 통해 학술적, 공중보건학적 측면의 성과를 기대할 수 있다. 먼저 학술적 측면에서는 구축된 데이터베이스 LCBB-SC를 통해 효율적인 정보 접근 및 검색을 비롯하여 만성 질환 관련 연구자들에게 연구의 기초가 되는 자료를 마련하는데 도움이 될 수 있을 것이다. 꾸준한 업데이트를 통해 더욱 방대한 양의 데이터를 축적하여 정확도를 높이고 유전적 데이터를 추가하여 이를 활용한 유전체학 및 바이오인포매틱스 연구 등 2차 연구의 기초 자료로 활용될 수 있을 것이다. 또한 현재 국내 만성질환 연구 분야에서는 잘 활용되고있지 않은 머신 러닝 및 프로그래밍을 활용한 만성 질환 예측 및 건강 지표 기준을 수립하여 프로그램화 하는 것에 도움이 될 것으로 기대된다. 향후 신뢰할 만한 수준의 데이터 종류와 양이 축적된다면 이를 의료 시스템에 도입시켜 환자를 진료하기 전 사전 점검 시스템으로도 활용 가능할 것이다. 뿐만 아니라, 만성질환 관련 정책 및 사전 예방을 위한 국가적 대책 마련에 대한 과학적 기반을 지원함으로써 국내 만성질환 발병률과 유병률을 낮추어 사회 및 경제적으로 발생되는 비용을 감소시키는데 도움이 될 수 있을 것으로 예상된다. 공중보건학적 측면에서는 웹 인터페이스를 통해 시스템의 접근성과 대중성을 높여 연구자들뿐만 아니라 일반인들에게 자가 증상 점검을 통해 본인의 건강 행태나 증상을 통해 발병 가능한 질환 목록을 예측률과 함께 제시하여 사전 예방에 도움이 될 수 있다. 또한 관련 질환에 대한 전문 정보를

제공하여 효율적인 정보 검색 및 접근이 가능하기 때문에 특히 예방이 중요한 만성 질환에 대한 사전 지식을 습득하여 국내 만성 질환 발병과 유병률을 낮추고 공중보건 향상에도 기여할 수 있을 것으로 기대된다.

# 제 5 장. 결론 및 총론

## 5.1 결 론

과학과 의학의 급진적인 발전으로 과거에는 불치병으로 분류되었던 많은 질병들이 현재는 백신이나 치료제에 의해 예방과 완치가 가능해 졌다. 과거에는 높은 사망률을 기록했던 전염병이나 감염병의 경우 현재는 백신의 보급으로 대부분 예방이 가능하며, 특히 미생물에 의한 감염은 종두법과 항생제의 개발로 더 이상 인간의 목숨을 위협하는 일이 드물게 되었다. 하지만 건강 행태나 식습관 의해 발생되는 질병의 경우 아직 뚜렷한 해결책이 제시되고 있지 않은 실정이다. 특히 만성질환의 경우 전 세계적으로 기대 수명이 높아짐과 동시에 유병률 또한 길어지게 되면서 관리가 잘 이루어지지 않을 시 이로 인한 합병증에 의해 현재 전 세계적으로 가장 많은 사망률을 기록하고 있다. 따라서 사전에 본인의 건강 상태를 주기적으로 파악하고 발병 가능한 질환들에 대한 신뢰할 만한 전문지식을 습득하여 예방이 최우선시 되어야 한다. 더욱이 현대인들의 삶은 갈수록 바빠지고 보살핌이 필요한 고령화 인구의 증가로 비교적 가벼운 질병들을 방치하게 됨으로써 심각한 질병이 되거나 단일 질병의 결과로 다발성 질환을 갖는 환자수가 크게 증가하고 있는 추세이다. 인터넷과 각종 미디어의 발달로 다양한 정보와 전문지식을 쉽게 접할 수 있어 현대인들의 삶의 질은 높아졌지만 생활 흐름 또한 급속도로 증가하게 되면서 건강을 돌보는 일에 소홀해지게 되었으며, 산업화와 도시화로 인해 지리적이나 경제적인 환경에 의해 의료 혜택을 받지 못하는 노인층이 증가하게 되면서 검증되지 않은

민간요법을 사용하거나 쉽게 치료가 가능한 질병 임에도 불구하고 적절한 치료시기를 놓쳐 질병을 악화시켜 결국 사망에 이르게 되는 경우가 증가하고 있다. 따라서 이러한 시간적, 지리적, 경제적 비용에서의 부담을 줄이고 본인의 건강 상태를 미리 파악하고 예방이 가능한 지능형 자가 진단 시스템을 통해 합병증으로 인한 피해를 최소화하고 좀 더 효율적인 질병 관리가 이루어져야 할 것이다. 본 연구에서는 이러한 필요성에 따라 국내 주요 사망 원인이 되는 질병들을 분류하고 대용량의 설문조사 자료와 참고 문헌 및 전문 서적을 통해 건강 지표가 될 수 있는 요인들에 각각의 가중치 값을 부여하여 최근 각광받고 있는 기계학습법에 적용시켜 연구를 실행하였다. 이를 체계적으로 데이터베이스화 시켜 관련분야 연구자들에게는 각 질환에 관련된 통합적인 데이터를 제공하여 연구에 이용 가능한 인프라를 제공하고자 하였으며, 전문 지식이 부족한 일반인들에게는 웹 인터페이스를 통해 시공간의 제약에서 벗어나 본인들의 건강 상태를 체크하고 간단한 증상들과 평소 생활 습관에 대한 세부적인 항목들을 선택하게 함으로써 발병 가능한 질병 목록과 이를 사전에 예방 가능한 지식 정보를 함께 제시하여 국민 보건에 기여하고자 하였다. 먼저, 통계청에서 제공하는 자료를 참고하여 국내 주요 사망 원인이 되는 질병들을 분류하고, 국민건강영양조사에서 제공하는 설문 조사 자료에서 주요 질병들의 지표로 활용될 수 있는 건강 행태 문답과 건강 상태 및 기존 의사의 진단으로 유병 여부를 파악할 수 있는 항목들을 선별하여 개인 아이디 별로 분류하여 질환 관련 통합 데이터베이스인 LCBB-SC를 구축하였다. 구축된 데이터베이스를 기반으로 본인이 자각할 수 있는 증상이나 직·간접적으로 질병 요인이 될 수 있는 건강 행태 목록을 대상으로 상관관계 분석과 머신 러닝 분석을 실행하여 사용된 알고리즘의 성능을 평가하고, 피드포워드 기법을 적용하여 데이터 양이 증가할수록 좀

더 정확한 분석이 가능할 수 있도록 시스템을 구축하였다. 각 질병에 주요 요인이 될 수 있는 지표들의 선정과정에서는 머신 러닝 알고리즘뿐 아니라 SAS를 이용한 상관 관계 분석과 distance matrix (DM)를 활용하여 검증 과정을 거쳤다. 그 결과 머신 러닝을 통해 분류된 질병들과 요인들 사이에 통계적으로도 높은 상관 관계가 있음을 확인할 수 있었다(p-value<0.0001). 향후 더 많은 데이터가 축적될 시 알고리즘의 특성상 정확하고 예측도가 높은 결과값을 추출할 수 있으며 이러한 메커니즘이 자가 진단 시스템에 유용하게 활용될 수 있음을 확인하였다. 추후 최신 연구 결과와 전문 정보를 추가하고 성능이나 속도 면에서 향상된 알고리즘으로 업데이트하면서 신뢰할 만한 시스템을 구축하여, 각종 질환들의 유병률과 발병률을 낮추고 위험 요소가 될 수 있는 잠재적인 건강 행태 개선을 통해 사전에 질병을 예방하여 질병 발생 피해를 최소화 시킴으로써 국민 건강 증진에 기여할 수 있을 것이라 기대한다.

## 5.2 총 론

고령화 시대에 접어들면서 지속적인 관리가 필요한 만성 질환자의 수가 전 세계적으로 꾸준히 증가하고 있으며, 대부분 여러 가지 질병을 동시에 가지고 있는 경우가 많아 이들 동반 질환에 대한 포괄적인 접근이 필수적이다. 하지만 현재 국내 주요 발병 질환 관리 체계와 연구는 과거 단일 질병을 중심으로 이루어진 것으로, 질환에 대한 예방적 관리 및 자가 관리가 점차 중요해지고 있는 현재의 역학적 변화를 반영하고 있지 못하고 있다. 따라서 질환의 효과적인 관리를 위해서는 기존의 치료 질병 중심의 접근 방식에서 예방 중심으로 확장하여 개인의 건강 상태를 미리 점검하고, 발병 가능한 질환에 대한 예방 지식을 습득하여 복합적인 다발성 질환으로 발생할 수 있는 각종 합병증 위험도를 파악하고 평가하는 것이 중요하다. 특히 만성질환의 경우는 건강 행태와 식습관에 크게 영향을 받는 질병이므로 해외의 연구 사례들과 웹사이트의 이용에는 많은 한계점이 존재한다. 따라서 본 연구에서는 정보통신 기술을 활용하여 국내에 특화된 만성 질환 관리 시스템 구축을 통해 국내 주요 사망원인이 되는 각종 질환 관련 요인을 파악하고 이를 제시함으로써 발병률과 유병률을 낮추고 건강 수명과 공중 보건 증진에 기여할 수 있는 토대를 마련하고자 하였다.

첫째, 다년간 축적된 국민 건강영양조사 설문조사 데이터를 바탕으로 국내 주요 발병 질환과 관련 요인들을 파악하고 이 데이터만을 추출하여 질병에 특화된 데이터베이스인 LCBB-SC를 구축하였다. 구축된 데이터베이스를 통해 일반인 및 연구자들의 정보 접근성을 높이고 관련 연구에 참고 자료로 사용 가능한 인프라를 제공하고자 하였다. 둘째, 구축한 데이터베이스를

활용하여 각종 질환의 위험 요소들과의 관련 및 연관도를 파악하기 위해 머신 러닝 및 통계적 분석을 시행하였다. 기존에 잘 알려진 질병 유발 요인들 외에도 잠재적 가능성이 있는 목록들을 산출하여 새로운 요인 분석 연구의 기초 자료가 될 수 있는 정보를 제공하고자 하였다. 예를 들어, 갑상선의 경우 과거에 알려진 요인 외에도 고혈압이나 심장질환과 비교적 높은 수치의 상관성이 추정되는 결과를 확인할 수 있었다. 이 외에도 가능성이 있는 다양한 요인들을 찾아내고자 하였으며 이를 수치와 함께 제시하였기 때문에 질병의 지표가 될 수 있는 새로운 요인을 찾는 연구에 적극 활용될 수 있을 것이다. 뿐만 아니라, 인공 신경망 알고리즘을 적용하여 다양한 유형의 데이터를 처리하고 이를 범주화 된 출력으로 통합시킨 결과를 제공하고 있다. 셋째, 웹 인터페이스를 통해 이용자로 하여금 간단한 기본 정보 및 개인의 건강 지표가 될 수 있는 평소의 생활 습관과 같은 요인에 따른 발병 가능 질환 목록을 예측률과 함께 제시해 주고 관련 질환 예방에 유용한 정보를 함께 제공함으로써 질병 유발에 직·간접적으로 가능성이 있는 건강 행태에 변화를 주어 직접적인 사전 관리를 통해 질병의 발병률을 낮추는데 기여할 수 있을 것이다. 또한 문헌 조사를 통해 얻은 관련 질환 전문적인 정보를 링크와 함께 제공함으로써 해당 질병에 대한 불필요한 검색 과정 없이 질병 지식과 예방 상식과 같은 정보 습득이 가능하도록 하였다.

해당 질환 관련 국내 연구자들에게는 약 3만명 가량의 질병 경험을 바탕으로 한 대용량의 설문 데이터베이스 시스템을 제공하고 대용량의 건강 지표 관련 데이터 연구에 활용 가능한 분석 기법을 함께 제시하여 보건학 분야 연구자들이 유용하게 이용할 수 있도록 인프라 마련하고, 최근 임상 실험 결과 및

질병관련 논문과 문헌 자료들을 적용시킨 업데이트 과정을 통해 최신 연구 동향에 활용될 수 있도록 하였다. 구축된 시스템은 국문과 영문으로 제공 되어 국내 연구자들뿐만 아니라 외국인 연구자들도 이용이 가능하며, 신경망 알고리즘을 적용하여 다양한 유형의 데이터를 처리하고 이를 범주화 된 출력으로 통합시킨 결과를 제공하고 있다. 순차적인 머신 러닝 학습 단계를 거쳐 모델의 파라미터 값을 조정하는 작업을 통해 적절한 모델을 찾아 예측률의 정확도를 높이고자 하였으나 설문 조사 자료의 경우 실제 임상 데이터에 비해 정확도나 신뢰도가 떨어진다는 단점이 있다. 따라서 이를 보완하기 위해 분석에 사용되는 데이터의 양을 늘리고, 유전적 요인이 크게 작용하는 만성 질환이나 돌연변이에 의한 악성 신생물의 경우 관련 질병 유전자 염기 서열이나 돌연변이 서열과 같은 유전적 데이터를 추가적인 분석 과정을 통해 향후 만성 질환에 대한 심도 있는 연구가 필요하다고 제시하는 바이다.

# 참고 문헌

Abeel T, Peer YV, Saeys Y. Java-ml: A machine learning library. Journal of Machine Learning Research. 2009;10(Apr):931-4.

Adams HP, Adams RJ, Brott T, Del Zoppo GJ, Furlan A, Goldstein LB, Grubb RL, Higashida R, Kidwell C, Kwiatkowski TG, Marler JR. Guidelines for the early management of patients with ischemic stroke: a scientific statement from the Stroke Council of the American Stroke Association. Stroke. 2003 Apr 1;34(4):1056-83.

Adams PF, Kirzinger WK, Martinez M. Summary health statistics for the US population: National Health Interview Survey, 2012. Vital and health statistics. Series 10, Data from the National Health Survey. 2013 Dec(259):1-95.

Adeli H, Hung SL. Machine learning: neural networks, genetic algorithms, and fuzzy systems. John Wiley & Sons, Inc.; 1994 Sep 30.

Akazaki K, Stemmermann GN. Comparative study of latent carcinoma of the prostate among Japanese in Japan and Hawaii. Journal of the National Cancer Institute. 1973 May 1;50(5):1137-44.

Alberti KG, Zimmet PF. Definition, diagnosis and classification of diabetes mellitus and its complications. Part 1: diagnosis and classification of diabetes mellitus. Provisional report of a WHO consultation. Diabetic medicine. 1998 Jul 1;15(7):539-53.

Alpaydin E. Introduction to machine learning. MIT press; 2014 Aug 22.

American Diabetes Association. Diagnosis and classification of diabetes mellitus. Diabetes care. 2010 Jan;33(Suppl 1):S62.

Bakris GL, Williams M, Dworkin L, Elliott WJ, Epstein M, Toto R, Tuttle K, Douglas J, Hsueh W, Sowers J. Preserving renal function in adults with hypertension and diabetes: a consensus approach. American journal of kidney diseases. 2000 Sep 1;36(3):646-61.

Barnett K, Mercer SW, Norbury M, Watt G, Wyke S, Guthrie B. Epidemiology of multimorbidity and implications for health care, research, and medical education: a cross-sectional study. The Lancet. 2012 Jul 7;380(9836):37-43.

Bayen RH, Davidson DJ, Bates DM. Mixed-effects modeling with crossed random effects for subjects and items. Journal of memory and language. 2008 Nov 1;59(4):390-412.

Besaratinia A, Pfeifer GP. Second-hand smoke and human lung cancer. The lancet oncology. 2008 Jul 1;9(7):657-66.

Blackburn BE, Ganz PA, Rowe K, Snyder J, Wan Y, Deshmukh V, Newman M, Fraser A, Smith K, Herget K, Kim J. Aging-Related Disease Risks among Young Thyroid Cancer Survivors. Cancer Epidemiology and Prevention Biomarkers. 2017 Nov 22.

Bloedorn E, Mani I, MacMillan TR. Machine learning of user profiles: Representational issues. arXiv preprint cmp-lg/9712002. 1997 Dec 9.

Boffetta P, Hashibe M, La Vecchia C, Zatonski W, Rehm J. The burden of cancer attributable to alcohol drinking. international Journal of Cancer. 2006 Aug 15;119(4):884-7.

Boffetta P, Hashibe M. Alcohol and cancer. The lancet oncology. 2006 Feb 1;7(2):149-56.

Botrugno I, Lovisetto F, Cobianchi L, Zonta S, Klersy C, Vailati A, Dionigi P, Jemos V. Incidental carcinoma in multinodular goiter: risk factors. The American surgeon. 2011 Nov 1;77(11):1553-8.

Brook DW, Brook JS, Zhang C, Cohen P, Whiteman M. Drug use and the risk of major depressive disorder, alcohol dependence, and substance use disorders. Archives of general psychiatry. 2002 Nov 1;59(11):1039-44.

Caballero JC, Martínez FJ, Hervás C, Gutiérrez PA. Sensitivity versus accuracy in multiclass problems using memetic pareto evolutionary neural networks. IEEE Transactions on Neural Networks. 2010 May;21(5):750-70.

Carretero OA, Oparil S. Essential hypertension: part I: definition and etiology. Circulation. 2000 Jan 25;101(3):329-35.

Casanovas O, Hicklin DJ, Bergers G, Hanahan D. Drug resistance by evasion of antiangiogenic targeting of VEGF signaling in late-stage pancreatic islet tumors. Cancer cell. 2005 Oct 31;8(4):299-309.

Cawley GC, Talbot NL. Efficient leave-one-out cross-validation of kernel fisher discriminant classifiers. Pattern Recognition. 2003 Nov 1;36(11):2585-92.

Chan PK, Stolfo SJ. Learning with non-uniform class and cost distributions: Effects

and a distributed multi-classifier approach. InIn Workshop Notes KDD-98 Workshop on Distributed Data Mining 1998.

Chan PK, Stolfo SJ. Toward Scalable Learning with Non-Uniform Class and Cost Distributions: A Case Study in Credit Card Fraud Detection. InKDD 1998 Aug 27 (Vol. 98, pp. 164-168).

Chen L, Aihara K. Chaotic simulated annealing by a neural network model with transient chaos. arXiv preprint chao-dyn/9701021. 1997 Jan 22.

Chesebro MJ. Passive smoking. American family physician. 1988 May;37(5):212-8.

Chobanian AV, Bakris GL, Black HR, Cushman WC, Green LA, Izzo JL, Jones DW, Materson BJ, Oparil S, Wright JT, Roccella EJ. Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure. hypertension. 2003 Dec 1;42(6):1206-52.

Chow WH, Swanson CA, Lissowska J, Groves FD, Sobin LH, Nasierowska-Guttmejer A, Radziszewski J, Regula J, Hsing AW, Jagannatha S, Zatonski W. Risk of stomach cancer in relation to consumption of cigarettes, alcohol, tea and coffee in Warsaw, Poland. International journal of cancer. 1999 Jun 11;81(6):871-6.

Chronic Kidney Disease Prognosis Consortium. Association of estimated glomerular filtration rate and albuminuria with all-cause and cardiovascular mortality in general population cohorts: a collaborative meta-analysis. The Lancet. 2010 Jun 12;375(9731):2073-81.

Cicėnas S, Vencevičius V. Lung cancer in patients with tuberculosis. World journal of surgical oncology. 2007 Dec;5(1):22.

Collobert R, Kavukcuoglu K, Farabet C. Torch7: A matlab-like environment for machine learning. InBigLearn, NIPS workshop 2011 (No. EPFL-CONF-192376).

Cosentino F, Volpe M. Hypertension, stroke, and endothelium. Current hypertension reports. 2005 Jan 1;7(1):68-71.

Diamantidis NA, Karlis D, Giakoumakis EA. Unsupervised stratification of cross-validation for accuracy estimation. Artificial Intelligence. 2000 Jan 1;116(1-2):1-6.

Dorland WA. Dorland's Illustrated Medical Dictionary32: Dorland's Illustrated Medical Dictionary. Elsevier Health Sciences; 2011.

Family doctor, http://familydoctor.org/familydoctor/en/health-tools/ search-by-symptom.html.

Farmer SE, Bernardotto M, Singh V. How good is Internet self-diagnosis of ENT symptoms using Boots WebMD symptom checker?. Clinical Otolaryngology. 2011 Oct 1;36(5):517-8.

Fontham ET, Correa P, Reynolds P, Wu-Williams A, Buffler PA, Greenberg RS, et al. Environmental tobacco smoke and lung cancer in nonsmoking women. A multicenter study. JAMA 1994;27:1752-9

Fortin M, Bravo G, Hudon C, Vanasse A, Lapointe L. Prevalence of multimorbidity among adults seen in family practice. The Annals of Family Medicine. 2005 May 1;3(3):223-8.

Frank MZ, Goyal VK. Capital structure decisions: which factors are reliably important?. Financial management. 2009 Mar 1;38(1):1-37.

Gardner MW, Dorling SR. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmospheric environment. 1998 Aug 1;32(14-15):2627-36.

Gelber RP, Gaziano JM, Manson JE, Buring JE, Sesso HD. A prospective study of body mass index and the risk of developing hypertension in men. American journal of hypertension. 2007 Apr 1;20(4):370-7.

Gordon Jr RS. An operational classification of disease prevention. Public health reports. 1983 Mar;98(2):107.

Govindarajan P, Ravichandran KS, Sundararajan S, Sreeja S. Impact of modifiable and non-modifiable risk factors on the prediction of stroke disease. InTrends in Electronics and Informatics (ICEI), 2017 International Conference on 2017 May 11 (pp. 985-989). IEEE.

Greene FL, Balch CM, Fleming ID, Fritz A, Haller DG, Morrow M, Page DL, editors. AJCC cancer staging handbook: TNM classification of malignant tumors. Springer Science & Business Media; 2002 May 10.

Gvero T, Kuncak V. Synthesizing Java expressions from free-form queries. InAcm Sigplan Notices 2015 Oct 23 (Vol. 50, No. 10, pp. 416-432). ACM.

Hall JE, do Carmo JM, da Silva AA, Wang Z, Hall ME. Obesity-induced hypertension: interaction of neurohumoral and renal mechanisms. Circulation research. 2015 Mar 13;116(6):991-1006.

Hasin DS, Goodwin RD, Stinson FS, Grant BF. Epidemiology of major depressive disorder: results from the National Epidemiologic Survey on Alcoholism and Related

Conditions. Archives of general psychiatry. 2005 Oct 1;62(10):1097-106.

Healthy Children symptom checker, http://www.healthychildren.org/english/tips-tools/symptom-checker.

Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. science. 2006 Jul 28;313(5786):504-7.

Hylek EM, Singer DE. Risk factors for intracranial hemorrhage in outpatients taking warfarin. Annals of internal medicine. 1994 Jun 1;120(11):897-902.

Isabel Symptom Checker, http://symptomchecker.isabelhealthcare. com/home/default

iTriage Symptom Checker, https://www.itriagehealth.com/

Kaluza B. Machine Learning in Java. Packt Publishing Ltd; 2016 Apr 29.

Kearns M, Ron D. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. Neural computation. 1999 Aug 15;11(6):1427-53.

Kim E, Choi JY, Lee KE. Management of thyroid nodules and cancers arising in the elderly. Journal of Korean Thyroid Association. 2012 Nov 1;5(2):99-103.

Klausen K, Borch-Johnsen K, Feldt-Rasmussen B, Jensen G, Clausen P, Scharling H, Appleyard M, Jensen JS. Very low levels of microalbuminuria are associated with increased risk of coronary heart disease and death independently of renal function, hypertension, and diabetes. Circulation. 2004 Jul 6;110(1):32-5.

Kontis V, Bennett JE, Mathers CD, Li G, Foreman K, Ezzati M. Future life expectancy in 35 industrialised countries: projections with a Bayesian model ensemble. The Lancet. 2017 Apr 1;389(10076):1323-35.

Kwon YJ, Kim SE, Park BJ, Bae JW, Kang HT. High-risk drinking is associated with dyslipidemia in a different way, based on the 2010–2012 KNHANES. Clinica Chimica Acta. 2016 May 1;456:170-5.

Levey AS, Coresh J. Chronic kidney disease. The lancet. 2012 Jan 14;379(9811):165-80.

Li X, Chen Y, Chen S, Sun G, Sun Y. The research of the association between dyslipidemia and chronic kidney disease. Biomedical Research. 2017;28(18).

Lindström J, Ilanne-Parikka P, Peltonen M, Aunola S, Eriksson JG, Hemiö K, Hämäläinen H, Härkönen P, Keinänen-Kiukaanniemi S, Laakso M, Louheranta A. Sustained reduction in the incidence of type 2 diabetes by lifestyle intervention:

follow-up of the Finnish Diabetes Prevention Study. The Lancet. 2006 Nov 11;368(9548):1673-9.

Luppino FS, de Wit LM, Bouvy PF, Stijnen T, Cuijpers P, Penninx BW, Zitman FG. Overweight, obesity, and depression: a systematic review and meta-analysis of longitudinal studies. Archives of general psychiatry. 2010 Mar 1;67(3):220-9.

Lyubimov D, Palumbo A. Apache Mahout: Beyond MapReduce. CreateSpace Independent Publishing Platform; 2016 Feb 18.

Lyznicki JM, Altman RD, Williams MA. Report of the American Medical Association (AMA) Council on Scientific Affairs and AMA recommendations to medical professional staff on the use of wireless radio-frequency equipment in hospitals. Biomedical instrumentation & technology. 2001;35(3):189-95.

Ma K, Baloch Z, He TT, Xia X. Alcohol consumption and gastric cancer risk: A meta-analysis. Medical science monitor: international medical journal of experimental and clinical research. 2017;23:238.

Mannino DM, Thorn D, Swensen A, Holguin F. Prevalence and outcomes of diabetes, hypertension and cardiovascular disease in COPD. European Respiratory Journal. 2008 Oct 1;32(4):962-9.

Marsland S. Machine learning: an algorithmic perspective. CRC press; 2015 Sep 15.

Mayo Clinic Symptom Checker, http://www.mayoclinic.com/health/symptom-checker/DS00671

McCallum AK. Mallet: A machine learning for language toolkit.

Memisevic R. Deep learning: Architectures, algorithms, applications. InHot Chips 27 Symposium (HCS), 2015 IEEE 2015 Aug 22 (pp. 1-127). IEEE.

Miura, Y., M. Saito, K. E. Sada, and N. Yajima. "AB0547 The prevalence and the risk factor of hypertension and dyslipidemia in systematic lupus erythematosus patients: exploratory research." (2017): 1242-1242.

Mouradian MS, Majumdar SR, Senthilselvan A, Khan K, Shuaib A. How well are hypertension, hyperlipidemia, diabetes, and smoking managed after a stroke or transient ischemic attack?. Stroke. 2002 Jun 1;33(6):1656-9.

Moussavi S, Chatterji S, Verdes E, Tandon A, Patel V, Ustun B. Depression, chronic diseases, and decrements in health: results from the World Health Surveys. The Lancet. 2007 Sep 8;370(9590):851-8.

Nahas AM, Bello AK. Chronic kidney disease: the global challenge. The Lancet. 2005 Jan 22;365(9456):331-40.

Ng andu NH. An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model. Statistics in medicine. 1997 Mar 30;16(6):611-26.

Nodhturft V, Schneider JM, Hebert P, Bradham DD, Russo K, ARNPC C, Aldahondo A, Clark V. Chronic disease self-management. Nurs Clin North Am. 2000 Jun;35:507-18.

Oh SW. Effects of Alcohol on Obesity and Metabolic Syndrome. 대한비만학회지.18(1): (2009).

Parlos AG, Chong KT, Atiya AF. Application of the recurrent multilayer perceptron in modeling complex process dynamics. IEEE Transactions on Neural Networks. 1994 Mar;5(2):255-66.

Pauwels RA, Buist AS, Calverley PM, Jenkins CR, Hurd SS. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: NHLBI/WHO Global Initiative for Chronic Obstructive Lung Disease (GOLD) Workshop summary. American journal of respiratory and critical care medicine. 2001 Apr 1;163(5):1256-76.

Pisano ED, Gatsonis C, Hendrick E, Yaffe M, Baum JK, Acharyya S, Conant EF, Fajardo LL, Bassett L, D'orsi C, Jong R. Diagnostic performance of digital versus film mammography for breast-cancer screening. New England Journal of Medicine. 2005 Oct 27;353(17):1773-83.

Polosa R, Morjaria JB, Caponnetto P, Battaglia E, Russo C, Ciampi C, Adams G, Bruno CM. Blood pressure control in smokers with arterial hypertension who switched to electronic cigarettes. International journal of environmental research and public health. 2016 Nov 11;13(11):1123.

Radzikowska E, Głaz P, Roszkowski K. Lung cancer in women: age, smoking, histology, performance status, stage, initial treatment and survival. Population-based study of 20 561 cases. Annals of oncology. 2002 Jul 1;13(7):1087-93.

Ramachandran A, Snehalatha C, Mary S, Mukesh B, Bhaskar AD, Vijay V. The Indian Diabetes Prevention Programme shows that lifestyle modification and metformin prevent type 2 diabetes in Asian Indian subjects with impaired glucose tolerance (IDPP-1). Diabetologia. 2006 Feb 1;49(2):289-97.

Ramana RV, Krishna B, Kumar SR, Pandey NG. Monthly rainfall prediction using

wavelet neural network analysis. Water resources management. 2013 Aug 1;27(10):3697-711.

Rimm EB, Williams P, Fosher K, Criqui M, Stampfer MJ. Moderate alcohol intake and lower risk of coronary heart disease: meta-analysis of effects on lipids and haemostatic factors. Bmj. 1999 Dec 11;319(7224):1523-8.

Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review. 1958 Nov;65(6):386.

Sacco RL, Kasner SE, Broderick JP, Caplan LR, Culebras A, Elkind MS, George MG, Hamdan AD, Higashida RT, Hoh BL, Janis LS. An updated definition of stroke for the 21st century: a statement for healthcare professionals from the American Heart Association/American Stroke Association. Stroke. 2013 Jul 1;44(7):2064-89.

Samet JM, Avila-Tang E, Boffetta P, Hannan LM, Olivo-Marston S, Thun MJ, Rudin CM. Lung cancer in never smokers: clinical epidemiology and environmental risk factors. Clinical Cancer Research. 2009 Sep 15;15(18):5626-45.

Savolainen MJ, Kesaniemi YA. Effects of alcohol on lipoproteins in relation to coronary heart disease. Curr Opin Lipidol 1995;6:243-50.

Schror K. Blood vessel wall interactions in diabetes. Diabetes. 1997 Sep 1;46:S115.

Silverman BL, Rizzo TA, Cho NH, Metzger BE. Long-term effects of the intrauterine environment: the Northwestern University Diabetes in Pregnancy Center. Diabetes care. 1998 Aug 1;21:B142.

Smith AD, Bolam JP. The neural network of the basal ganglia as revealed by the study of synaptic connections of identified neurones. Trends in neurosciences. 1990 Jul 1;13(7):259-65.

Sowers JR, Epstein M, Frohlich ED. Diabetes, hypertension, and cardiovascular disease: an update. Hypertension. 2001 Apr 1;37(4):1053-9.

St Sauver JL, Boyd CM, Grossardt BR, Bobo WV, Rutten LJ, Roger VL, Ebbert JO, Therneau TM, Yawn BP, Rocca WA. Risk of developing multimorbidity across all ages in an historical cohort study: differences by sex and ethnicity. BMJ open. 2015 Feb 1;5(2):e006413.

Stamler J, Vaccaro O, Neaton JD, Wentworth D, Multiple Risk Factor Intervention Trial Research Group. Diabetes, other risk factors, and 12-yr cardiovascular mortality for men screened in the Multiple Risk Factor Intervention Trial. Diabetes care. 1993 Feb 1;16(2):434-44.

Stuifbergen AK, Rogers S. Health promotion: An essential component of rehabilitation for persons with chronic disabling conditions. Advances in Nursing Science. 1997 Jun 1;19(4):1-20.

Sugomori Y. Java Deep Learning Essentials. Packt Publishing Ltd; 2016 May 30.

Thun MJ, Hannan LM, Adams-Campbell LL, Boffetta P, Buring JE, Feskanich D, Flanders WD, Jee SH, Katanoda K, Kolonel LN, Lee IM. Lung cancer occurrence in never-smokers: an analysis of 13 cohorts and 22 cancer registry studies. PLoS medicine. 2008 Sep 9;5(9):e185.

Toniato A, Bernardi C, Piotto A, Rubello D, Pelizzo MR. Features of papillary thyroid carcinoma in patients older than 75 years. Updates in Surgery. 2011 Jun 1;63(2):115-8.

Tsai CF, Hsu YF, Lin CY, Lin WY. Intrusion detection by machine learning: A review. Expert Systems with Applications. 2009 Dec 1;36(10):11994-2000.

US Environmental Protection Agency. Respiratory health effect of passive smoking: lung cancer and other disorders. (EPA/600/6-90/006F). Washington DC: US EPA, Office of Research and Development RD-689, 1992

Valderas JM, Starfield B, Sibbald B, Salisbury C, Roland M. Defining comorbidity: implications for understanding health and health services. The Annals of Family Medicine. 2009 Jul 1;7(4):357-63.

Vestbo J, Hurd SS, Agustí AG, Jones PW, Vogelmeier C, Anzueto A, Barnes PJ, Fabbri LM, Martinez FJ, Nishimura M, Stockley RA. Global strategy for the diagnosis, management, and prevention of chronic obstructive pulmonary disease: GOLD executive summary. American journal of respiratory and critical care medicine. 2013 Feb 15;187(4):347-65.

Vogeli C, Shields AE, Lee TA, Gibson TB, Marder WD, Weiss KB, Blumenthal D. Multiple chronic conditions: prevalence, health consequences, and implications for quality, care management, and costs. Journal of general internal medicine. 2007 Dec 1;22(3):391-5.

Wagena EJ, Knipschild PG, Huibers MJ, Wouters EF, Van Schayck CP. Efficacy of bupropion and nortriptyline for smoking cessation among people at risk for or with chronic obstructive pulmonary disease. Archives of Internal Medicine. 2005 Oct 24;165(19):2286-92.

Wakabayashi I, Kobaba-Wakabayashi R, Masuda H. Relation of drinking alcohol to

atherosclerotic risk in type 2 diabetes. Diabetes Care. 2002 Jul 1;25(7):1223-8.
Wakabayashi I, Kobaba-Wakabayashi R, Masuda H. Relation of drinking alcohol to atherosclerotic risk in type 2 diabetes. Diabetes Care. 2002 Jul 1;25(7):1223-8.

WebMD, http://symptoms.webmd.com/

WHO, http://www.who.int/en/

World Health Organization (WHO). World health statistics 2015. World Health Organization; 2015 May 14.

Xie X, Atkins E, Lv J, Bennett A, Neal B, Ninomiya T, Woodward M, MacMahon S, Turnbull F, Hillis GS, Chalmers J. Effects of intensive blood pressure lowering on cardiovascular and renal outcomes: updated systematic review and meta-analysis. The Lancet. 2016 Jan 30;387(10017):435-43.

Xin X, Wang W, Xu H, Li Z, Zhang D. Exposure to Chinese famine in early life and the risk of dyslipidemia in adulthood. European journal of nutrition. 2018 Jan 3:1-8.

Yazdi FT, Clee SM, Meyre D. Obesity genetics in mouse and human: back and forth, and back again. PeerJ. 2015 Mar 24;3:e856.

Yu S, Sun Y. GW26-e3861 Prevalence of dyslipidemia and associated factors among the hypertensive population from rural Northeast China: an update. Journal of the American College of Cardiology. 2015 Oct 20;66(16 Supplement):C115.

Zaharia M, Xin RS, Wendell P, Das T, Armbrust M, Dave A, Meng X, Rosen J, Venkataraman S, Franklin MJ, Ghodsi A. Apache spark: a unified engine for big data processing. Communications of the ACM. 2016 Oct 28;59(11):56-65.

Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. Journal of biomedical informatics. 2002 Oct 1;35(5-6):352-9.

Karayiannis N, Venetsanopoulos AN. Artificial neural networks: learning algorithms, performance evaluation, and applications. Springer Science & Business Media; 2013 Jun 29.