사물인터넷융복합논문지 Vol. 11, No. 5, pp. 157-162, 2025 https://doi.org/10.20465/KIOTS.2025.11.5.157

전자건강기록(EHR)의 프라이버시 보호를 위한 효율적 동형암호 기법 제안

장우혁[1], 이근호[2*]
[1]백석대학교 컴퓨터공학부 학생, [2]백석대학교 컴퓨터공학부 교수

An Efficient Homomorphic Encryption Scheme for Privacy-Preserving Electronic Health Records (EHRs)

Woo-Hyuk Jang[1], Keun-Ho Lee[2*]
[1]Student, Division of Computer Engineering, Baek-Seok University
[2]Professor, Division of Computer Engineering, Baek-Seok University

요 약 전자건강기록(EHR)은 환자의 진료 정보를 포함한 의료 데이터로 공중보건 및 임상 연구에서 중요한 가치를 지닌다. 그러나 개인정보보호법, GDPR(General Data Protection Regulation), HIPAA(Health Insurance Portability and Accountability Act) 등 규제로 인해 연구자는 원본 데이터에 직접 접근할 수 없어 활용에 제약이 따른다. 기존 연구는 동형암호(Homomorphic Encryption, HE)를 적용해 집단 수준 분석을 가능하게 했으나 높은 연산량으로 실용성이 제한되었다. 본 연구는 이를 해결하기 위해 모집군에서 실제 등장한 질병 코드만 추출하고, 제로패딩으로 동일 길이 벡터를 구성한 뒤 슬롯 순서 무작위화(Random Slot Permutation, RSP) 방식을 적용하는 기법을 제안한다. 이로써 HE 연산 호환성을 확보하면서 통계적 패턴 노출을 차단하고, 연구자는 매핑 슬롯을 활용해 원하는 통계치를 효율적으로 집계할 수 있다. 제안 기법은 프라이버시를 보장하면서도 단순하고 실용적인 집단 데이터 분석을 지원한다.

주제어 : 전자건강기록, 동형암호, 프라이버시 보호, 제로패딩, 랜덤 슬롯 무작위화

Abstract Electronic health records (EHRs) are medical data including patients' medical information and have important value in public health and clinical research. However, due to regulations such as the Personal Information Protection Act, GDPR, and HIPAA, researchers cannot directly access the original data, limiting their use. Existing studies have applied homogeneous encryption (HE) to enable group-level analysis, but their practicality has been limited due to the high amount of computation. To solve this problem, this study proposes a technique that extracts only the disease code that actually appeared in the recruitment group, constructs the same length vector with zero padding, and applies the Random Slot Permutation (RSP) method. This blocks statistical pattern exposure while ensuring HE operation compatibility, and enables researchers to efficiently aggregate desired statistics using mapping slots. The proposed technique ensures privacy while also supporting simple and practical collective data analysis.

Key Words : Electronic Health Records, Homomorphic Encryption, Privacy, Zero Padding, Random Slot Permutation

이 논문은 2025학년도 백석대학교 학술연구비 지원을 받아 작성되었음
*교신저자 : 이근호(root1004@bu.ac.kr)
접수일 2025년 10월 10일 수정일 2025년 10월 15일 심사완료일 2025년 10월 17일

1. 서론

전자건강기록(Electronic Health Records, EHR)은 환자의 진료 이력, 질병 코드, 검사 수치 등을 포함하는 핵심 의료 데이터로, 공중보건과 임상 연구에서 중요한 통계적 가치를 지닌다. 대규모 EHR 데이터는 질병 발병률, 치료 효과, 위험 요인 분석 등 다양한 의학적 통찰을 제공해 의료 혁신과 정책 수립에 활용된다. 그러나 해당 데이터는 민감한 개인정보를 포함하고 있어 개인정보보호법, GDPR, HIPAA 등에 따라 엄격히 관리되며[1, 2], 연구자가 개별 환자 데이터에 직접 접근하기는 어렵다[3, 4].

기존 공개키 기반 암호화 기법은 저장 및 전송 과정의 기밀성을 보장했으나, 내부자 공격이나 권한 오남용과 같은 보안 위협은 여전히 존재하였다. 또한 전통적 암호화는 암호화된 상태에서 연산을 지원하지 못해 연구자가 직접 통계 분석을 수행하기에는 제약이 있었다. 이를 보완하기 위해 동형암호(Homomorphic Encryption, HE)가 주목받고 있으며[5], HE는 암호화된 상태에서도 덧셈과 곱셈 연산을 지원해 연구자가 민감한 정보를 열람하지 않고도 집단 수준의 평균이나 비율을 계산할 수 있다[6]. 그러나 높은 연산 복잡도로 인해 대규모 EHR 데이터에 적용하기에는 실용성이 떨어진다.

환자별 진료 기록은 수많은 질병 코드 중 일부만 포함해 다수의 0이 발생하는 희소 벡터(Sparse Vector) 구조를 가진다[7]. 본 연구는 이러한 특성을 활용하여 모집군(Cohort)에서 실제 등장한 질병 코드만 추출하고, HE의 연산 호환성을 확보하기 위해 제로 패딩 해 동일 길이 벡터를 구성한 뒤 슬롯 순서 무작위화(Random Slot Permutation, RSP)를 적용하는 방법을 제안한다. 이를 통해 모든 환자 데이터가 동일한 구조에서 연산할 수 있으며, 통계적 패턴 노출을 차단하면서 효율적인 집단 통계 분석을 지원한다.

2. 관련연구

2.1 기존 암호화 기법

현재 EHR 보호에는 AES, RSA, ECC와 같은 전통적 암호화 기법이 널리 사용되고 있다. 이러한 방식은 데이터 저장 및 전송 과정에서 기밀성을 보장하는 데 효과적이지만, 내부자 공격이나 접근 권한 오남용에는 취약하다[8, 9]. 또한 전통적 암호화는 암호화된 상태에서 연산을 수행할 수 없으므로, 연구자가 직접 데이터를 이용하여 통계 분석을 수행하는 데에는 근본적인 한계가 존재한다.

2.2 동형 암호(HE)

HE는 암호화된 상태에서 연산을 수행할 수 있는 기법으로, 민감한 데이터를 보호하면서도 통계적 분석을 가능하게 한다. 특히 EHR과 같이 직접 열람이 제한되는 의료 데이터의 활용에 효과적이다. 연구자는 개인정보를 침해하지 않고도 집단 수준의 평균이나 비율을 계산할 수 있으며, 이는 GDPR, HIPAA 등 엄격한 규제 환경에서도 합법적이고 안전한 데이터 활용을 가능하게 한다.

HE의 기본 원리는 다음과 같다. 두 평문 m_1, m_2가 있을 때, 이들을 각각 암호화한 후 암호문 연산 덧셈($\oplus$)을 수행한 뒤 복호화하면 다음과 같은 결과를 얻을 수 있다.

$$Dec(Enc(m_1) \oplus Enc(m_2)) = m_1 + m_2$$

즉, 암호문 상태에서 덧셈을 수행하면 복호화 결과는 평문 덧셈과 일치한다. 따라서 여러 데이터의 합을 안전하게 계산할 수 있으며, 이를 다시 $\frac{1}{n}$ (n : 표본 수)과 곱함으로써 평균 역시 산출할 수 있다[10, 11].

2.3 HE 최적화 연구

HE는 암호화된 상태에서 연산을 수행 가능하다는 장점에도 불구하고, 기본적으로 매우 높은 연산량을 요구한다. 이에 따라 실제 응용 환경에서의 사용성을 높이기 위해 다양한 최적화 기법이 제안되었다. 예컨대 CSR (Compressed Sparse Row)과 같은 희소 행렬 표현을 암호화 처리에 적용하거나, 불필요한 0 값을 제거하는 방식은 저장 공간과 연산 부담을 줄이는 데 효과적이다. 최근에는 사용자–아이템 행렬을 CSR 표현과 CKKS (Cheon-Kim-Kim-Song) 기반 FHE로 통합하여 암호화 도메인에서 직접 처리하고, 통신 및 연산 비용을 절감하는 방법도 제안되었다[12].

이러한 방식은 일반적인 수치 행렬에는 효과적이지만, EHR처럼 질병 코드의 존재 여부를 중심으로 하는 데이터 구조와는 차이가 있다. 단순 압축만으로는 EHR의 통계적 특성과 질의 요구를 충분히 반영하기 어렵고, 연구자가 특정 질병의 유병률이나 집단 통계를 계산하려면 여전히 별도의 마스킹 연산이 필요하다. 이는 HE 환경에

서 성능 저하로 이어질 수 있으며, 따라서 의료 데이터의 구조적 특성을 고려한 별도의 최적화 접근이 요구된다.

2.4 희소 표현(Sparse Representation)

희소 표현(Sparse Representation, SR)은 데이터 내 다수의 0 값을 저장하지 않고 실제 값이 존재하는 원소와 해당 인덱스만 저장하는 방식이다. 이를 통해 대규모 데이터의 저장 공간 요구와 연산 부담을 줄일 수 있다. 반면 일반적인 밀집 표현(Dense Representation)은 모든 요소를 그대로 저장하기 때문에 불필요한 0 값까지 포함되어 비효율적이다. SR은 이러한 한계를 보완하여 압축률과 연산 효율을 개선하며, 특히 EHR에서와 같이 대부분의 질병 코드가 0으로 나타나는 데이터 구조에 적합하다.

3. 제안 방법

3.1 설계 개요

본 연구는 EHR의 희소성과 HE의 연산 제약을 동시에 고려하여, RSP와 제로 패딩(Zero Padding)을 결합한 기법을 제안한다. 모집군 내에서 실제로 등장한 질병 코드를 추출한 뒤, 제로 패딩을 통해 모든 환자 벡터의 길이를 동일하게 맞추고 각 환자 데이터에 동일한 슬롯 구조를 부여한다.

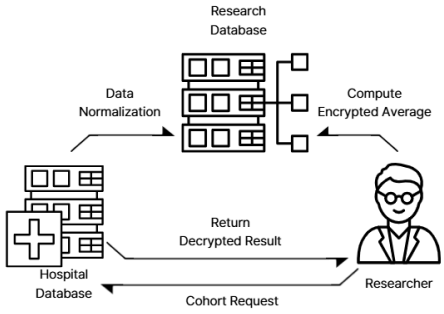
이후 슬롯의 순서를 무작위로 재배치함으로써 특정 질병의 데이터 분포가 외부로 유추되거나 노출되는 위험을 차단한다. 이러한 재배치 과정은 환자별 벡터가 무작위로 혼합되도록 하여, 구조적 동일성과 통계적 은닉성을 동시에 확보한다[13].

이러한 구조는 HE 환경에서 구조적 혼합성을 보장하며, 연구자가 별도로 관리되는 매핑 슬롯 정보를 활용해 원하는 질병 통계를 단순한 마스킹 연산만으로 효율적 집계가 가능하다.

본 연구의 접근법은 기존의 압축 기반 방식과 달리, 데이터 분포 은닉과 질의 처리 효율성을 동시에 달성할 수 있다는 점에서 차별성을 갖는다.

3.2 구조도

[Fig. 1]은 본 연구에서 제안하는 시스템 구조를 나타낸다. 시스템은 크게 병원 데이터베이스, 연구용 데이터베이스, 연구자 측 질의 처리 과정으로 구성된다.



[Fig. 1] Proposed system architecture

연구자가 특정 모집군을 요청하면, 병원 데이터베이스는 해당 환자의 EHR을 검색하고, 이를 희소 벡터 형태로 변환한다. 이후 모집군 전체 질병 코드 집합 개수를 기준으로 제로 패딩을 적용하여 모든 환자 벡터의 길이를 동일하게 정규화한다.

정규화된 벡터는 RSP를 통해 슬롯 순서가 무작위로 재배치되며, 생성된 매핑 정보는 암호문과 분리되어 안전하게 관리된다. 이후 데이터는 HE 방식으로 암호화되어 연구용 데이터베이스에 저장된다.

연구자가 질의를 수행하면, 연구용 데이터베이스는 매핑 슬롯 정보를 활용하여 관련 슬롯만 선택적으로 식별하고, 나머지 슬롯은 마스킹 처리한 후 암호화된 상태에서 평균 계산 연산을 수행한다. 계산된 암호화 결과는 병원 데이터베이스로 전달되어 복호화되며, 최종 통계 결과만 연구자에게 반환된다.

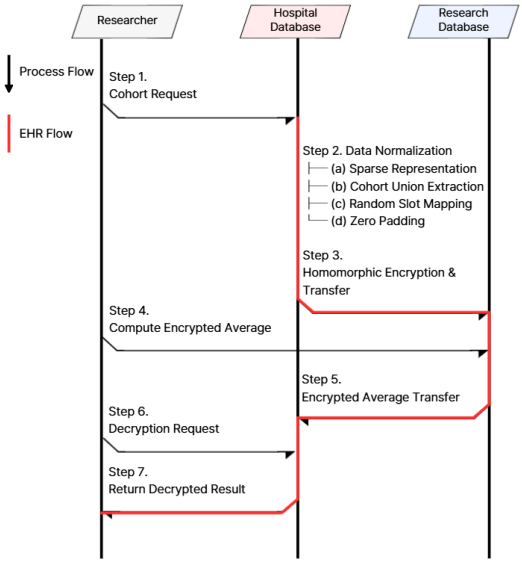
이러한 구조는 개별 환자의 민감한 데이터를 직접 노출하지 않으면서도 모집군 단위의 통계 분석을 가능하게 한다. 또한 ICD-10 전체 코드(약 7만 개)를 직접 암호화하는 기존 방식과 비교할 때, 실제로 등장한 질병 코드만을 대상으로 연산이 수행되므로 암호화 비용이 대폭 감소하고 처리 효율성이 향상된다.

3.3 동작 절차

[Fig. 2]는 제안 기법의 동작 절차를 단계별로 요약한 것이다.

Step 1. 연구자가 특정 모집군을 요청하면, 병원 데이터베이스는 해당 환자의 EHR을 검색하고 원본 데이터 대신 벡터 형태의 표현으로 변환한다.

Step 2. 데이터 정규화(Data Normalization)는 다음 네 단계로 구성된다.



[Fig. 2] Workflow of the proposed system

(a) 각 환자의 진료 이력을 희소 벡터 형태로 표현하는 Sparse Representation을 수행한다.

(b) 모집군 내에서 실제 등장한 질병 코드의 합집합을 추출하는 Cohort Union Extraction을 수행한다.

(c) 각 벡터의 슬롯을 무작위로 재배치하기 위해 RSP를 적용한다.

(d) 마지막으로 제로 패딩을 통해 모든 환자 벡터의 길이를 동일하게 맞춘다.

Step 3. 정규화 및 매핑이 완료된 데이터는 HE 방식으로 암호화되어 연구용 데이터베이스로 전송된다.

Step 4. 연구자는 암호화된 상태에서 통계적 분석(예: 평균 계산)을 요청하고, 연구용 데이터베이스는 이에 대응하는 암호문 연산을 수행한다.

Step 5. 계산된 암호화 결과(Encrypted Average)는 병원 데이터베이스로 전송된다.

Step 6. 병원은 결과 암호문을 복호화하여 최종 통계 결과를 생성한다.

Step 7. 복호화된 결과(Decrypted Result)는 연구자에게 반환된다.

이 절차는 연구자가 민감한 개인 데이터에 접근하지 않고도 모집군 단위의 통계치를 얻을 수 있게 하며, 병원은 데이터 통제권을 유지하면서 프라이버시 보호와 분석 효율성을 동시에 달성할 수 있다.

4. 실험 결과

4.1 실험 환경

본 연구의 실험은 Windows 11 운영체제에서 WSL2(Windows Subsystem for Linux 2)를 활용하여 Ubuntu 기반 가상화 환경을 구동한 상태에서 수행하였다. 암호화 연산은 Python 환경에서 Pyfhel 라이브러리를 이용하여 구현하였으며, 암호 스킴으로는 CKKS 방식을 채택하였다.

데이터셋은 실제 EHR을 대체하기 위하여 ICD-10 체계를 참고하여 총 70,000개의 진단 코드를 가상 생성하였다.

환자 수는 1,000명으로 설정하였으며, 각 환자가 보유하는 질병 코드의 수는 1개에서 최대 10개 사이로 할당하였다. 또한, 질병 코드의 분포는 현실적 편중 현상(일부 질환의 집중 발생)을 반영하기 위해 Zipf 분포(s=1.15)를 따르도록 설계하였다. 이를 통해 제안된 기법이 대규모이면서도 편향성을 지닌 의료 데이터 환경에서도 효과적으로 적용될 수 있음을 검증하였다.

4.2 비교 기준

본 연구에서는 Dense 방식(70,000개의 ICD-10 전체 코드를 모두 암호화하는 전통적 접근)과 제안 기법을 동일한 CKKS 파라미터 환경에서 비교하였다. 평가의 초점은 암호화 비용에 두었으며, 이는 환자 수와 데이터 크기에 따라 전체 처리 성능에 중요한 영향을 끼칠 뿐 아니라 저장 및 전송 비용과도 밀접하게 연관되기 때문이다.

비교를 위해 세 가지 지표를 중심으로 성능을 정량적으로 분석하였다. 첫째, 환자당 평균 암호화 시간(T_{enc})을 측정하여 두 방식 간 처리 속도를 평가하였다. 둘째, 환자당 평균 암호문 개수(CT)를 산출하여 자원 소모를 비교하였다. 셋째, 환자당 평균 암호문 크기(S_{CT})를 측정하여 저장 공간 및 전송 효율성을 검증하였다.

$$T_{enc} = \frac{\text{Total Enc. Time}}{N},$$
$$CT = \frac{\text{Total Ciphertexts}}{N},$$
$$S_{CT} = \frac{\text{Total Ciphertext Size}}{N}$$

이러한 지표들은 실제 의료 데이터 환경에서 연구자와 기관 간 데이터 공유 과정에서 초래될 수 있는 연산 지연, 저장 부담, 네트워크 전송 비용을 평가하는 핵심 기준으로 활용된다[14, 15].

4.3 성능 평가

〈Table 1〉은 Dense 방식과 제안 기법의 성능 비교 결과를 나타낸다. Dense 방식은 70,000개의 ICD-10 전체 코드를 모두 암호화하기 때문에 환자당 평균 9개의 암호문이 생성되었으며, 총 암호화 시간은 9.01초가 소요되었다. 반면, 제안 기법은 모집단에서 실제로 등장한 코드만을 반영함으로써 환자당 단일 암호문으로 집약할 수 있었고, 전체 코드 차원을 235개로 축소하였다. 이에 따라 암호화 시간 또한 0.86초로 대폭 단축되었으며, 암호문 크기 또한 약 7.5MB에서 0.84MB로 감소하였다.

〈Table 1〉 Performance comparison of encryption

Name	T_{enc}	CT	S_{CT}
Dense	9.01 s	9	7.5 MB
Proposed	0.86 s	1	0.84 MB

이러한 결과는 제안 기법이 HE 기반 의료 데이터 분석 환경에서 암호화 효율성을 현저히 향상시킬 수 있음을 입증하며, 특히 대규모 집단 연구에서 실용성을 크게 높일 수 있음을 시사한다.

5. 결론

본 논문에서는 EHR의 희소성과 HE의 연산 제약을 동시에 고려하여, RSP와 제로 패딩을 결합한 기법을 제안하였다.

제안 방식은 모집군 내 실제 등장한 질병 코드 집합을 기준으로 환자 데이터를 동일 길이 벡터로 정규화하고, 슬롯 위치를 무작위로 재배치하여 통계적 패턴 노출을 효과적으로 차단한다. 이를 통해 연구자는 원본 데이터에 직접 접근하지 않고도 모집군 단위 통계 결과를 획득할 수 있으며, 병원은 개인정보 보호를 유지하면서 효율적인 집계 처리를 수행할 수 있다.

본 연구의 의의는 단순한 저장 공간 절감이나 연산 효율 향상에 그치지 않고, 의료 데이터의 구조적 특성을 반영해 프라이버시 보장과 활용성 개선을 동시에 달성한 데 있다. 기존 암호화 방식은 집계 연산을 지원하지 못하고, HE는 높은 연산 비용으로 실용성이 제한적이었다. 제안 기법은 모집군 기반 정규화와 슬롯 무작위화를 통해 연산 호환성을 확보하고 데이터 분포 노출 위험을 최소화함으로써, 프라이버시 보호가 요구되는 의료 협력 연구나 공공보건 통계 분석 환경에서 유용하게 적용될 수 있다.

그러나 본 연구는 몇 가지 한계를 내포한다. 우선, 논의가 주로 설계와 가능성에 집중되어 있어 실제 의료 데이터셋을 대상으로 한 실험적 검증은 아직 수행되지 않았다. 또한 제안 기법은 모집군 내 실제 등장한 질병 코드 집합을 기준으로 정규화를 수행하기 때문에, 분석 대상 모집군이 달라질 경우 병원 데이터베이스에 재요청이 필요하다는 점에서 운영상 비효율이 발생할 수 있다.

이러한 문제점을 보완하기 위해 향후 연구에서는 동적 확장 매핑 전략을 적용하여 모집군 변화에도 기존 암호문을 재활용할 수 있는 방안을 모색할 예정이다. 더 나아가 실제 임상 데이터 환경을 기반으로 한 실험적 검증과 다양한 분석 시나리오 확장 연구를 통해 제안 기법의 안전성과 실용성을 한층 강화할 계획이다.

REFERENCES

[1] J.Jonnagaddala and Z.S.-Y.Wong, "Privacy-preserving strategies for electronic health records in the era of large language models," npj Digital Medicine, Vol.8, No.1, Art.34, 2025,.

[2] A.K.Conduah, S.Ofoe and D.Siaw-Marfo, "Data privacy in healthcare: Global challenges and solutions," Digital Health, Vol.11, Art.20552076251343959, 2025

[3] C.H.Lee, K.H.Lim and S.Eswaran, "A comprehensive survey on secure healthcare data processing with homomorphic encryption: attacks and defenses," Discover Public Health, Vol.22, No.1, p.137, 2025.

[4] J.Scheibner, M.Ienca and E.Vayena, "Health data privacy through homomorphic encryption and distributed ledger computing: an ethical-legal qualitative expert assessment study," BMC Medical Ethics, Vol.23, No.1, p.121, 2022.

[5] O.G.d'Aliberti and M.A.Clark, "Preserving patient privacy during computation over shared electronic health record data," Journal of Medical Systems, Vol.46, No.12, 2022

[6] K.Munjal and R.Bhatia, "A systematic review of homomorphic encryption and its contributions in healthcare industry," Complex Intelligent Systems, pp.1-28, 2022.

[7] J.Kauffman, R.Miotto, E.Klang, A.Costa, B.Norgeot, M.Zitnik, S.Khader, F.Wang, G.N.Nadkarni and B.S.Glicksberg, "Embedding methods for electronic health record research," Annual Review of Biomedical Data Science, Vol.8, pp.563-590, 2025.

[8] I.Lee, "Analysis of insider threats in the healthcare

industry: a text mining approach," Information, Vol.13, No.9, 2022.

[9] W.Hurst, B.Tekinerdogan, T.Alskaif, A.Boddy and N.Shone, "Securing electronic health records against insider-threats: A supervised machine learning approach," Smart Health, Vol.26, p.100354, 2022.

[10] T.V.T.Doan, M.-L.Messai, G.Gavin and J.Darmont, "A survey on implementations of homomorphic encryption schemes," The Journal of Supercomputing, Vol.79, No.13, pp.15098-15139, 2023.

[11] A.Acar, H.Aksu, A.S.Uluagac and M.Conti, "A survey on homomorphic encryption schemes: theory and implementation," arXiv preprint arXiv:1704.03578, 2017.

[12] M.N.Chowdhury, A.Bauer and M.Zhou, "Efficient privacy-preserving recommendation on sparse data using fully homomorphic encryption," arXiv preprint arXiv:2509.03024, 2025.

[13] D.Kahrobaei, A.Wood and K.Najarian, "Homomorphic encryption for machine learning in medicine and bioinformatics," ACM Computing Surveys, 2020.

[14] L.Jiang and L.Ju, "FHEBench: Benchmarking fully homomorphic encryption schemes," arXiv preprint arXiv:2203.00728, 2022.

[15] V.Sidorov, E.Y.F.Wei and W.K.Ng, "Comprehensive performance analysis of homomorphic cryptosystems for practical data processing," arXiv preprint arXiv:2202.02960, 2022.

장 우 혁(Woo-Hyuk Jang) [준회원]

■ 2023년 3월 ~ 현재 : 백석대학교 컴퓨터공학부

〈관심분야〉
블록체인, 암호학, 인공지능, 시스템 보안

이 근 호(Keun-Ho Lee) [종신회원]

■ 2006년 8월 : 고려대학교 컴퓨터학과(이학박사)
■ 2006년 9월 ~ 2010년 2월 : 삼성전자 DMC연구소 기술전략팀 과장
■ 2010년 3월 ~ 현재 : 백석대학교 컴퓨터공학부 교수

〈관심분야〉
융합보안, 블록체인, 개인정보보호, 이동통신 보안