Journal of Korea Multimedia Society Vol. 25, No. 10, October 2022(pp. 1512-1523)
https://doi.org/10.9717/kmms.2022.25.10.1512

혈액암 인자 유효성 검증과 분류를 위한 진단 예측 알고리즘 성능 비교 분석

정재승[†], 주현수[††], 조치현[†††]

Comparative Analysis of Diagnostic Prediction Algorithm Performance for Blood Cancer Factor Validation and Classification

Jeong Jae-Seung[†], Hyunsu Ju[††], Chi-Hyun Cho[†††]

ABSTRACT

Artificial intelligence application in digital health care has been increasing with its development of artificial intelligence. The convergence of the healthcare industry and information and communication technology makes the diagnosis of diseases more simple and comprehensible. From the perspective of medical services, its practice as an initial test and a reference indicator may become widely applicable. Therefore, analyzing the factors that are the basis for existing diagnosis protocols also helps suggest directions using artificial intelligence beyond previous regression and statistical analyses. This paper conducts essential diagnostic prediction learning based on the analysis of blood cancer factors reported previously. Blood cancer diagnosis predictions based on artificial intelligence contribute to successfully achieve more than 90% accuracy and validation of blood cancer factors as an alternative auxiliary approach.

Key words: Artificial Intelligence, Digital Health Care, Machine Learning, Diagnostic Prediction Learning, Classifier

1. 서 론

최근 인공지능의 빠른 발전이 여러 분야에 적용되면서 다양하고 깊이 있는 인공지능 관련 연구 및 관심이 꾸준히 증대되어오고 있다. 특히, 인간의 개인 건강과 질환을 관리하는 의료분야에서 인공지능을 이용한 디지털 헬스케어 등 관련 연구 및 산업들이 발전해 오고 있다. 일반적인 스마트 워치와 같은 스마트 기기를 활용한 기초 건강에 관한 기술인 디지털 헬스케어 분야 외에도 특이 질병이나 고위험군 질병들에 대한 접근을 조금 더 편리하게 하는 관점의 연구 또한 많아지고 있다[1-4]. 이러한 연구들을 통해 다양한 인공지능 예측 알고리즘 활용으로부터 의료 서비스에 대한 경량화나 정확성을 향상하게 시킴으로써 개인의 건강한 삶과 의료 시스템의 질적인 향상을 기대할 수 있다[5]. 일반적인 진단 예측 알고리즘의 성능을 확인하기 위해서는 주로 공개된 데이터들 (WDBC data set, pima indian diabetes data set 등)

※ Corresponding Author : Hyunsu Ju, Address: 5, Hwarang-ro 14-gil Seongbuk-gu Seoul, 02792 Republic of Korea, TEL : +82-2-958-5362, E-mail : hyunsuju@kist.re.kr
Receipt date : Jul. 25, 2022, Revision date : Sep. 29, 2022
Approval date : Oct. 4, 2022
[†] Korea Institute of Science and Technology, Post-Silicon Semiconductor Institute
wotmd104@kist.re.kr
[††] Korea Institute of Science and Technology, Post-Silicon Semiconductor Institute
hyunsuju@kist.re.kr
[†††] Department of Laboratory Medicine, Korea University Ansan Hospital
9754091@korea.ac.kr
※ This research was supported by the Korean National Police Agency (KNPA)-(PR08-04-000-21).

을 사용한다. 이러한 공개 데이터를 기반으로 검증된 기본 기계학습 기법들을 바탕으로 실제 측정된 데이터들에 대한 인자들의 상관관계 분석과 유효성 검증은 기존의 통계 분석에 대한 보조 수단으로 활용할 수 있을 것이다. 본 논문에서는 통계 분석으로 이미 분석된 데이터에 대해 기계학습 기법을 적용함으로써 새로 연구된 인자들에 대한 진단예측의 활용과 관련한 유효성을 확인하려 한다.

본 논문에서는 최근 보고된 호중구 젤라티나제 관련 리포칼린(neutrophil gelatinase-associated lipocalin, NGAL) 인자와 혈액암 데이터의 분석 결과를 활용하여[6], 혈액암의 진단 예측 정확도를 바탕으로 NGAL 인자의 유효성을 검증할 것이다. 기존에 보고된 데이터 분석 및 질병 진단 예측에 활용 가능한 k-최근접 이웃(k-nearest neighbor)[7-9], 랜덤 포레스트(random forest)[10-12], 순수 베이즈 분류기(naive bayes' classifier)[13-17]의 다양한 기계학습 기법들을[18] 본 논문에서 분류기로써 사용하여 진단 예측을 진행할 것이다. NGAL은 다양한 원인으로부터 급성 신장 손상이 발생하면 급증하는 소변 생물학적 표지자로써 많은 연구가 이루어지고 있다[19]. 또한, NGAL은 급성신장 손상 외에도 염증과 연관성을 가지고 있다고 보고되어 있다[20]. 최근에는 만성 염증이 혈액암의 발달을 유도한다는 연구 가설이 제기되었으며[21-23], 염증 유발 관련 인자에 대한 분석을 통해 혈액암의 메커니즘을 밝히는 데 도움이 될 것이다. 골수(bone marrow, BM) NGAL 수준은 혈액암이 있는 개인에서 말초혈액(peripheral blood, PB) NGAL 수준보다 더 높은 것으로 보고되었으며[24], 이는 BM NGAL이 PB NGAL보다 BM 미세환경을 더 잘 반영할 것을 시사한다. 최근 연구에는 혈액암 환자의 BM NGAL 수준을 정상 BM의 NGAL 수준과 비교하여 호중구(neutrophil) 수를 포함한 혈액학적 측정 지표와 BM NGAL의 연관성이 제시되었다[6]. 혈액암과 같이 희귀 질병의 경우 데이터의 개수가 부족하기 때문에 제시된 연관성의 추가적인 유효성 확인이 필요하다. 이를 위해, 혈액학적 측정 지표와 BM NGAL 연관성에 대해 추가적인 유효성 확인을 위한 보조 기법으로써 인공지능 알고리즘을 적용하여 확인하였다. 이전 연구에서[6] 사용된 데이터를 활용하여 위에서 설명한 기계학습 기법들로부터 진단 예측 학습 및 이에 관한 결과를 비교하였다. 혈액암과 BM NGAL 수치에 대한 연관성을 확인하기 위해 다섯 가지 혈액암 환자와 일반인의 데이터를 분류하여 진단 예측을 확인하였으며, 진단 예측에서 사용되는 데이터의 측정 지표에 NGAL을 포함한 결과와 포함하지 않은 결과를 비교함으로써 BM NGAL 유효성을 확인하였다. 골수 측정 및 혈액암의 희귀성 때문에 측정 검체수가 106개로써 한계가 존재하지만 적은 데이터에도 높은 분류 정확도를 보이는 k-NN[25] 및 NBC를[26] 포함한 기본적인 진단 예측 분류기들로부터 혈액암 질병을 분류하는데 높은 예측 정확도 및 BM NGAL의 연관성을 확인할 수 있었고 기존의 통계 분석에 대한 의의를 구체화하였다. 또한, 사용된 분류기 종류별 예측진단 정확도 비교로부터 진단 예측에 활용된 데이터의 적합한 분류기를 특정하였으며 이를 통하여 추후 예측진단으로써의 적합한 분류기로써 GNBC의 활용가능성을 확인할 수 있었다.

2. 이 론

2.1 k-최근접 이웃(k-nearest neighbors algorithms, k-NN)

k-최근접 이웃 알고리즘은 1951년 Evelyn Fix와 Joseph Hodges에 의해 처음 개발되었으며[27] 후에, Thomas Cover에 의해 확장된 알고리즘이다[28]. 패턴인식 분야에서 회귀, 분류에 사용되는 비모수 지도학습 방법으로써 두 가지 경우 모두 입력이 특징 공간 내에서 k개의 가장 가까운 데이터들에 대해 분석하며 출력은 분류, 회귀의 목적에 따라 다르다. 본문에서 사용할 k-NN 분류에서의 출력값은 소속된 항목 값이다. 데이터들은 k개의 가장 가까운 이웃들 사이에서 가장 공통된 항목에 할당되고 과반수에 의해 분류된다. 가장 간단한 기계학습 알고리즘으로써 데이터 군집화(clustering)에 주로 사용된다. 이때 k 값에 따라 분류 결과값이 많은 영향을 줄 수 있기 때문에 지도학습을 통해 최적값 k에 대해 도출한다. k 값이 너무 작으면 데이터의 잡음에 의한 과대 적합 가능성이 있으며 반대로 k 값이 너무 크면 과소 적합이 될 수 있다. 그래서 최적 k에 대해서 지도학습을 수행했을 시에 테스트 데이터에 대해 과대, 과소 적합을 피하고자 적절한 k를 선택해야 한다. Fig. 1은 Hb와 NGAL에 대해 혈액암 두 가지 질병군과 일반군

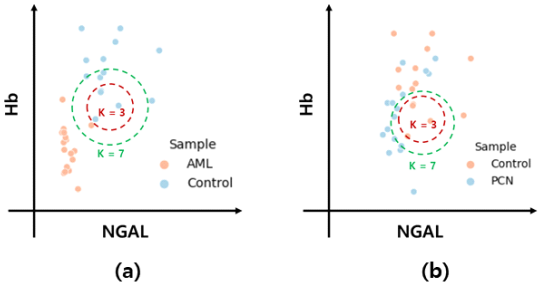

Fig. 1. Schematic illustration of k–NN classifications. (a) "AML vs. control group" classification and (b) "PCN vs. control group" classification for Hb and NGAL according to k values.

(control)을 분류하는 분류기로써 k-NN을 적용한 예시이다. 본 연구에서는 질병군 종류별로 k-NN에서의 k 값에 따른 분류 정확도를 비교하고 최적 k 값에 대한 k-NN을 적용하여 인자별 유효성 및 분류기별 정확도를 비교하였다. 측정 지표로써 NGAL의 유무에 따른 k-NN 분류결과가 유의미한 차이가 있는지 확인을 통해 k-NN을 사용한 혈액암 분류에서의 NGAL과 혈액암과의 연관성을 확인할 수 있다.

2.2 랜덤 포레스트(random forest, RF)

랜덤 포레스트 또는 랜덤 의사결정 포레스트(random decision forest)는 분류, 회귀 및 기타 작업을 위한 앙상블 학습 방법으로 데이터를 학습할 때 다수의 의사결정 트리를 구성하여 작동한다[29]. 분류기로써의 랜덤 포레스트의 출력은 의사결정 트리에서 선택한 클래스들에 대한 다수결 결과 값이다. 랜덤 포레스트를 구성하는 의사결정 트리는 여러 규칙을 순차 적용하여 독립 변수 공간을 분할한다. 먼저, 특정된 독립 변수에 대한 기준값을 분류 규칙으로써 정하고, 학습에 사용되는 전체 데이터 집합을 독립변수 기준값과 대소 비교를 통해 데이터 그룹을 나눈다. 데이터 그룹을 나눌 때 사용되는 기준값으로 지니 계수(gini coefficient)를 사용한다. 지니 계수는 통계학적으로 불균형의 정도를 표현하는 계수로써 불순도의 개념을 가지고 있으며 0에 가까울수록 불순도가 낮고 균등하게 분리되었음을 의미하며, 반대로 지니 계수가 1에 가까울수록 불순도가 높고 분리가 불균등함을 의미한다[30]. 마지막으로, 각각 나눠진 그룹에 대해 다시 이전 단계와 마찬가지의 기준 설정 및 대소 비교를 반복하여 트리구조를 나눠진 그룹의 클래스가 단일 클래스가 될 때까지 시행한다. 이러한 의사결정 트리를 다양한 독립 변수 기준값에 대해 랜덤성을 가지기 때문에 트리들이 서로 다른 특성을 가지며 이로부터 예측 동작에 대해 비상관화(decorrelation) 특성을 가짐으로써 일반화 성능을 향상시킨다.

2.3 순수 베이즈 분류기(naive bayes' classifier, NBC)

순수 베이즈 분류기는 베이즈 정리(bayes' theorem)를 기계 학습에 적용한 일종의 확률 분류기로써 분류할 데이터 클래스 간의 독립성을 가정한다[31]. 순수 베이즈 분류기의 장점은 다음과 같다. 첫째, 일부 확률 모델에서 순수 베이즈 분류는 지도 학습 환경에서 간단하고 빠르며 효율적으로 훈련될 수 있다. 둘째, 분류에 필요한 측정 지표를 추정하기 위한 훈련 데이터의 양이 매우 적다. 셋째, 단순한 베이지안 분류는 단순한 설계와 가정에도 불구하고 많은 복잡한 실제 상황에서 잘 작동한다. 순수 베이즈 분류기에서 분류할 인스턴스는 N개의 지표가 독립변수인 벡터 $\boldsymbol{d} = (d_1, \ldots, d_n)$으로 표현된다. 순수 베이즈 분류는 이 벡터를 사용하여 식 1과 같이 가능한 확률적 결과 k 클래스를 할당하여 조건부 확률에 대해 사후 확률 (post probability)을 계산하여 분류한다.

$$P(h_k|d_1, \cdots, d_n) \text{ , } (\boldsymbol{d} = d_1, d_2, \cdots, d_n) \quad (1)$$

$$P(h_k \mid \boldsymbol{d}) = \left(\frac{P(\boldsymbol{d} \mid h_k) P(h_k)}{P(\boldsymbol{d})} \right)$$

2.4 가우시안 순수 베이즈 분류기(Gaussian naive Bayes' classifier, GNBC)

순수 베이즈 분류기에 사용되는 데이터의 형태와 분포에 따라 이벤트 모델을 달리 설정하여 사용할 수 있다. 본 논문에서는 진단 예측을 위한 생체 데이터에 대한 분류를 진행하기 때문에 연속적인 데이터에 적용할 수 있는 가우시안 순수 베이즈 분류기를 사용한다[32]. GNBC를 적용하기 위해 분류하려는 데이터의 분포가 가우스 분포를 따른다고 가정하기 때문에 데이터를 구성하는 지표들에 대한 분포를 확인하고 측정 지표화 하여 정규분포식으로 표현하고 이에 대한 순수 베이즈 분류기에서의 사전확률과 가능도의 곱에 대한 연산을 통해 분류한다. Fig. 2.는 본 논문에서 사용한 GNBC의 구성 및 연산을 도식화

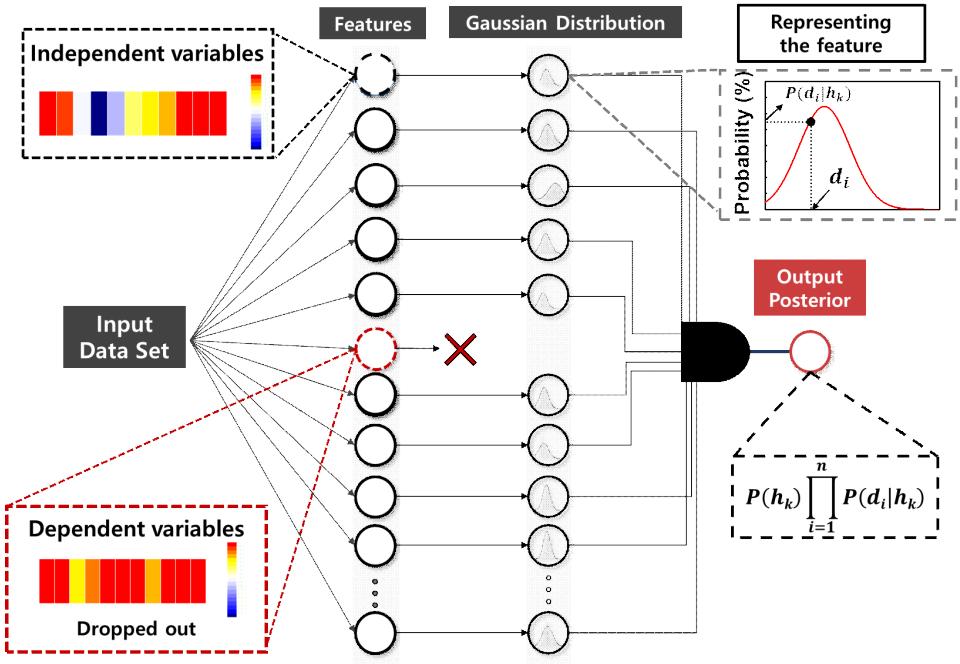

Fig. 2. Schematic diagram of GNBC configuration and classification process.

한 것이다. 데이터 지표들의 분포를 학습하고 이를 기반으로 테스트 데이터 지표 값으로부터 가능도를 구하고, 이를 통해 사후 확률을 계산하여 진단 예측을 수행한다. 혈액암군과 일반군에서의 데이터의 지표들의 분포 차이가 클수록 진단 예측 정확도가 높을 것이다.

2.5 골수 상청액 호중구 젤라티나제 관련 리포칼린 및 혈액학적 악성종양의 혈액학적 측정 지표 분석

최근 만성 염증이 혈액암의 발달로 이어진다는 가설이 제기되고 있다[24]. 이 가설에 따르면, 염증 유발자들에 대한 분석을 통해 혈액암의 메커니즘을 설명할 수 있다. 혈액암 환자의 호중구 젤라티나제 관련 리포칼린(neutrophil gelatinase-associated lipocalin, NGAL) 수준을 평가하는 이전 연구에서 혈액암을 가진 개인에서 인간 골수(bone marrow, BM) NGAL의 수치가 말초혈액(peripheral blood, PB) NGAL 보다 훨씬 더 높게 나타났으며, 이는 BM NGAL이 PB NGAL보다 BM 미세 환경을 더 잘 반영할 것임을 시사한다[33]. 혈액암과 NGAL 연관성 분석을 위해서는 PB NGAL뿐만 아니라 BM NGAL의 측정 또한 필요하다. 본 연구에서는 혈액암 환자들에 대한 검체 데이터들을 기계학습 기법을 이용해 분류하고, 혈액학적 양성 진단 분류기의 정확도에 대한 BM NGAL 수치의 기여도를 확인하였다.

3. 제안한 방법

3.1 진단 예측을 위한 데이터 분류 전처리

혈액학적 측정 지표 분석에 사용된 데이터를 사용하여 진단 예측을 진행하기 위해 고려대학교 진단검사의학과에서 측정된 106개의 검체 데이터를 사용하였다[6]. 검체 데이터는 급성 골수성 백혈병(acute myeloid leukemia, AML) 17개, 만성 골수성 백혈병(chronic myeloid leukemia, CML) 12개, 골수이형성증후군(myelodysplastic syndrome, MDS) 25개, 골수증식성 종양(myeloproliferative neoplasm, MPN) 22개, 형질세포 종양(plasma cell neoplasm, PCN) 16개의 진단 샘플 및 대조군 14개로 구성되어있다. 검체 데이터는 총 18개의 측정 지표로 구성되어 있으나 누락 혹은 이상 값을 포함한 지표는 진단 예측 학습 데이터에서 제외하여 13개의 측정 지표를 활용하였다. 13개의 측정 지표는 신체 나이와 BM NGAL 수치 및 헤모글로빈(hemoglobin, Hb), 백혈구 수(white

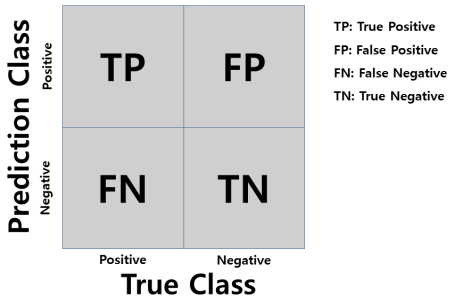

Fig. 3. Configuration of confusion matrix.

blood cell count, WBC count), 호중구 수(neutrophil count), 혈소판 수(platelet count), BM 성숙 골수 세포와 에리티로이드 비율(the ratio of maturing myeloid cells to erythroid cells in the bone marrow, M:E ratio), BM 세포 % (BM 흡인 슬라이드에 대한 BM 세포 수) 와 같은 혈액학적 측정 지표에 대하여 수집되었다. BM 세포 % 에는 BM 블라스트(BM blast), BM 프로미셀로사이트(BM promyelocyte), BM 골수세포성(BM cellularity), BM 메타미셀로사이트(BM metamyelocyte), BM 밴드 호중구(BM band neutrophil), BM 호중구(BM neutrophil)로 구성되어있다. 혈액암 종류별 진단에 대한 학습 및 검증 데이터의 비율은 0.75, 0.25 비율로 구성하였다.

3.2 연구 데이터 진단 예측을 통한 분류기 성능 비교

실제 측정된 검체 데이터를 이용하여 분류기들의 성능을 비교하였다. 전체 106개의 측정 데이터 중 4종류(AML, MDS, MPN, PCN)의 진단 샘플 데이터들을 각각 대조군과 분류하여 진단을 예측하는 성능을 확인함으로써 분류기 기법들의 성능을 비교 하였다. 또한, 측정 지표 중 BM NGAL의 수치와 혈액암의 관련성을 확인하기 위해, BM NGAL 측정 지표 유무를 조작 변인으로 설정하여 진단 예측 분류 성능을 확인하였다. 제한적인 데이터 개수의 한계점을 보완하기 위해 k-fold cross-validation을 적용하여 최종적으로 비교하였다.

3.3 진단 예측 분류기 비교를 위한 오차 행렬 결과 비교

오차 행렬(confusion matrix)은 학습된 분류 모델의 예측(prediction) 성능을 측정하기 위해서 모델이 얼마나 정답과 오답에 대해 혼동(confused) 하는지 쉽게 알 수 있도록 예측오류를 4개의 분면 (이진 분류 기준)에 요약하여 직관적으로 표현하는 표이다. Fig. 3은 오차 행렬의 구성이며, 표의 각 행은 예측 클래스의 인스턴스를 표현하고 각 열은 실제 클래스의 인스턴스를 표현한다. Fig. 3에서의 TP는 true positive로써 positive를 실제 positive로 참(true) 예측한 것을 의미하며, FP는 false positive로써 실제 데이터는 negative인데 positive로 거짓(false) 예측했음을 의미한다.

4. 실험 결과 및 고찰

4.1 검체 데이터 AML 진단 예측 분류기 비교

Fig. 4는 검체 데이터에 대한 AML 및 control 그룹에 대한 분류기별 분류 성능을 확인하기 위한 오차 행렬이다. BM NGAL의 여부와 상관없이 동일한 결과를 보임으로써 AML과 일반 control 그룹에 대한

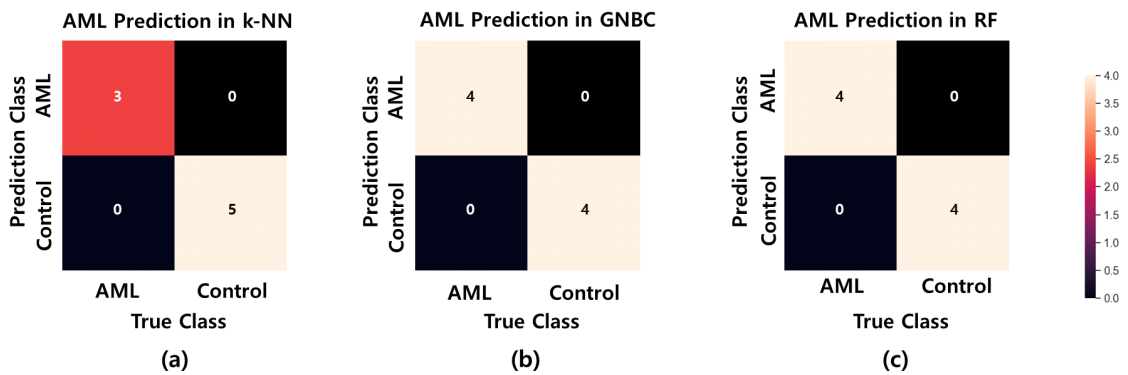

Fig. 4. Confusion matrices of AML diagnostic prediction cases according to different classifier types. (a) k-NN, (b) GNBC, and (c) RF.

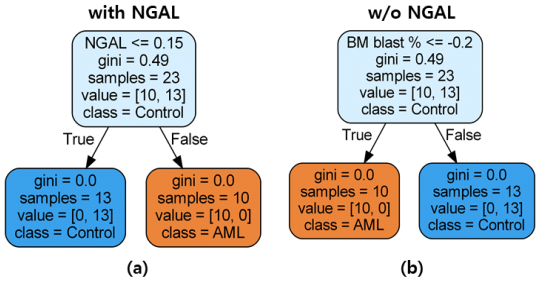

Fig. 5. Sample tree of RF used in AML diagnostic prediction (a) with NGAL feature and (b) without NGAL feature.

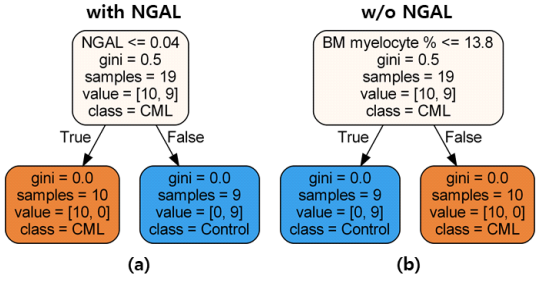

Fig. 7. Sample tree of RF used in CML diagnostic prediction (a) with NGAL feature and (b) without NGAL feature.

분류에 대해서는 세 종류의 분류기에서 모두 정확히 진단함을 확인할 수 있다.

Fig. 5는 RF를 사용한 AML 진단 예측에 대해 NGAL을 포함한 분류 (a)와 NGAL을 포함하지 않은 분류 (b)에 대한 결과를 sample tree 결과로써 비교하였다. 두 결과 모두 decision single tree의 깊이가 1 수준이며 26개 학습 데이터에 대해 명확히 구분 가능함이 확인되었다. NGAL은 분류 지표로써 AML을 예측 진단하는 데에 중요한 지표라 할 수 있지만, NGAL이 없는 경우에도 BM blast % 지표를 기준으로 1의 깊이만으로 데이터가 분류 되는 것을 확인하였다. 따라서, AML 그룹과 control 그룹을 분류할 때 NGAL이 중요 지표로 사용될 수 있으나 필수적인 지표는 아님을 확인할 수 있었다.

4.2 검체 데이터 CML 진단 예측 분류기 비교

Fig. 6은 검체 데이터에 대한 CML 및 control 그룹에 대한 분류기별 진단 예측 오차 행렬이다. CML에 대한 분류도 AML과 마찬가지로 BM NGAL 지표 여부와 상관없이 동일한 결과를 가진다. AML과 마찬가지로 테스트 그룹 데이터가 7개 밖에 되지 않는 한계는 존재하지만 분류기별 학습 및 분류가 높은 정확도를 보임을 확인하였다.

NGAL 유무와 상관없이 동일한 분류 정확도를 보여주는 CML 역시 Fig. 7에서도 마찬가지로 RF에서의 NGAL 외에 BM myelocyte % 가 중요 지표로써 의미를 가짐을 확인할 수 있었다.

4.3 검체 데이터 MDS, MPN 진단 예측 분류기 비교

검체 데이터 MDS와 MPN에 대한 진단 예측은 AML과 CML에 대한 진단 예측과 다르게 일부 분류기에서 오답으로 예측하는 결과를 보여준다. Fig. 8은 MDS에 대한 진단 예측 오차 행렬이다. Fig. 8 (c) RF 분류 결과에서는 10개 테스트 데이터에 대한 예측이 올바른 결과를 보이지만, Fig. 8 (a) k-NN, (b) GNBC 결과에서는 false positive에 대한 오류를

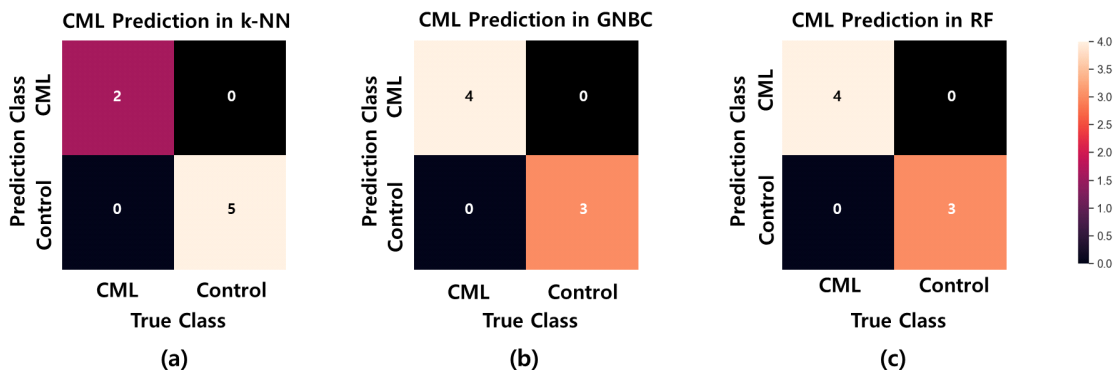

Fig. 6. Confusion matrices of CML diagnostic prediction cases according to different classifier types. (a) k-NN, (b) GNBC, and (c) RF.

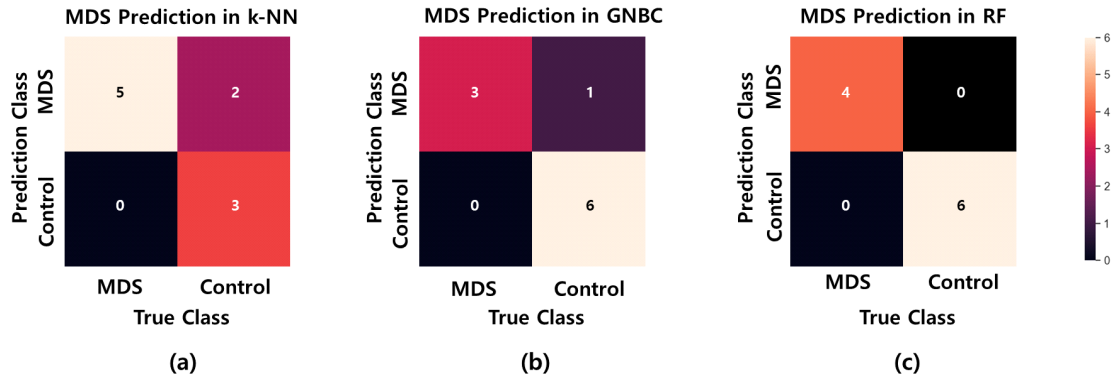

Fig. 8. Confusion matrices of MDS diagnostic prediction cases according to different classifier types. (a) k-NN, (b) GNBC, and (c) RF.

발생시킨다. 반대로 Fig. 9의 (a), (b)에서 보여주듯 MPN에 대한 진단 예측은 Fig. 9 (a) k-NN 및 (b) GNBC 결과에서는 높은 분류 정확도를 보이나, (c) RF 분류 결과에서는 마찬가지로 false positive에서 오차를 발생시킨다. MPN과 MDS의 데이터 지표 및 분류기에 따른 진단 예측이 차이점을 가짐을 알 수 있다.

4.4 검체 데이터 PCN 진단 예측 분류기 비교

Fig. 10은 검체 데이터 PCN에 대한 분류기별 진단 예측 오차행렬 결과이다. Fig. 10 (b) NBC, (c) RF에 결과에서는 제대로 된 진단 예측을 할 수 없음을 확인했으며 Fig. 10 (a) k-NN 분류기에서만 87.5%의 분류 정확도를 확인할 수 있었다. 또한, Fig. 11에서 보여주듯이 NGAL 지표의 유무에 따른 진단 예측 분류에서는 큰 차이를 보였으며 그 결과 PCN의 경우 NGAL 유무에 따른 k-NN에서의 진단 예측 정확도 차이로부터 NGAL 지표가 주요 PCN 데이터 분류에 유효할 수 있는 결과를 확인하였다.

4.5 NGAL 지표 유무 검체 데이터에 대한 분류기별 k-fold 교차검증 비교

Table 1 및 Table 2는 NGAL 지표 유무에 따른 검체 데이터와 분류기별 진단 예측 정확도를 보여준다. 제한적인 데이터 개수의 한계점을 보완하기 위해 k-fold cross validation을 진행하였으며, 다섯 종류의 혈액암 진단 예측 분류에서 각 실험군과 대조군의 샘플 개수 중 더 적은 샘플 개수를 k-fold의 최대 k값으로 설정하였다. Table 1과 2에서 보여주듯이 AML, CML, MDS는 평균적으로 95% 수준의 높은 정확도

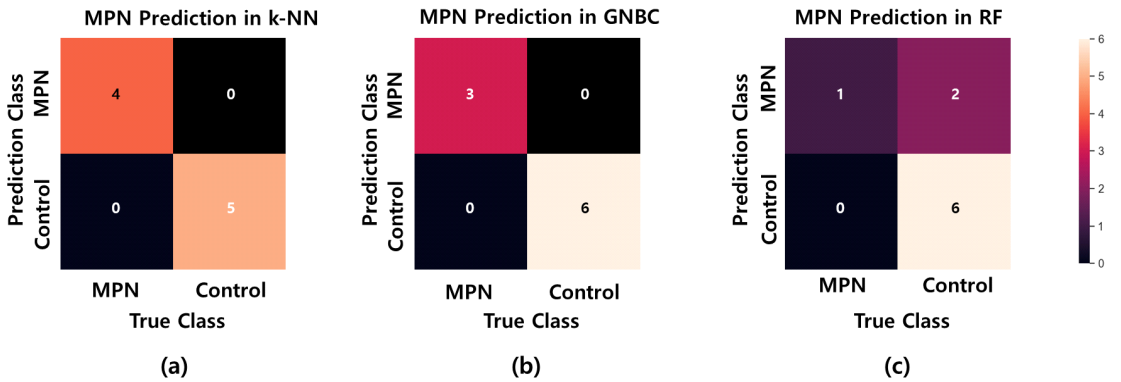

Fig. 9. Confusion matrices of MPN diagnostic prediction cases according to different classifier types. (a) k-NN, (b) NBC, and (c) RF.

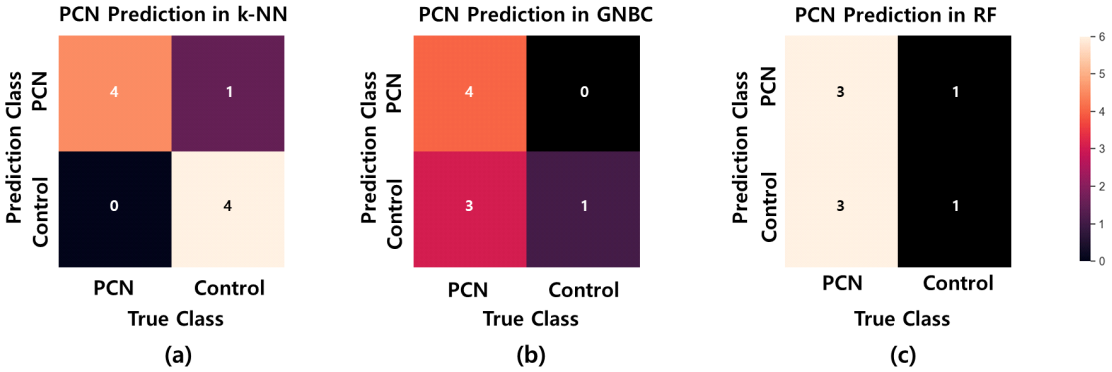

Fig. 10. Confusion matrices of PCN diagnostic prediction cases according to different classifier types. (a) k-NN, (b) GNBC, and (c) RF.

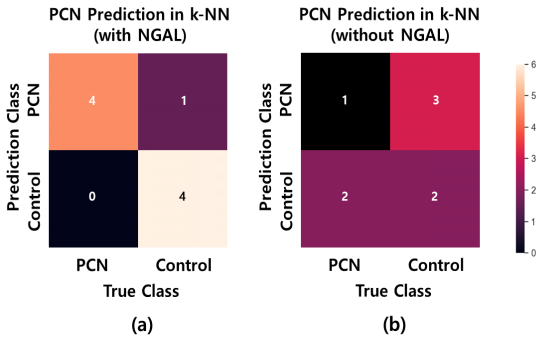

Fig. 11. Confusion matrices of PCN diagnostic prediction cases by k-NN (a) with NGAL and (b) without NGAL.

를 보여줬으나 MPN과 PCN에서의 진단 예측 정확도는 평균 73.9%로써 낮은 예측 정확도를 보인다. 또한, NGAL 유무에 따라 예측 정확도의 차이는 존재하지만, 분류기별 유의미한 정확도 차이는 확인할 수 없었다. 또한, Fig. 12에서 보여주듯이 RF 분류기에서의 PCN 진단 예측에 대한 데이터 특성 중요도를 확인하였을 때 NGAL의 중요도보다 M:E ratio, BM band neutrophil %, BM myelocyte % 의 중요도가 높음을 확인하였다. RF 분류기를 사용한 PCN에 대한 진단 예측에 있어서 NGAL은 연관성은 존재하나 절대적인 지표가 아님을 확인할 수 있었다.

Table 1. Diagnostic prediction accuracy of each classifier for blood cancer sample data with NGAL (k-fold cross-validation).

Classifier	Disease (with NGAL feature)				
	AML k=14	CML k=12	MPN k=14	MDS k=14	PCN k=14
	k-fold cross-validation average Accuracy [%]				
k-NN	100	100	78.6	83.3	70.2
GNBC	**100**	**100**	**84.5**	**96.4**	**66.6**
RF	96.4	95.8	66.7	92.9	75

Table 2. Diagnostic prediction accuracy of each classifier for blood cancer sample data without NGAL (k-fold cross-validation).

Classifier	Disease (without NGAL feature)				
	AML k=14	CML k=12	MPN k=14	MDS k=14	PCN k=14
	k-fold cross-validation average Accuracy [%]				
k-NN	91.7	100	78.6	75	73.8
GNBC	**100**	**100**	**83.3**	**100**	**66.7**
RF	96.4	83.3	73.8	94.0	69

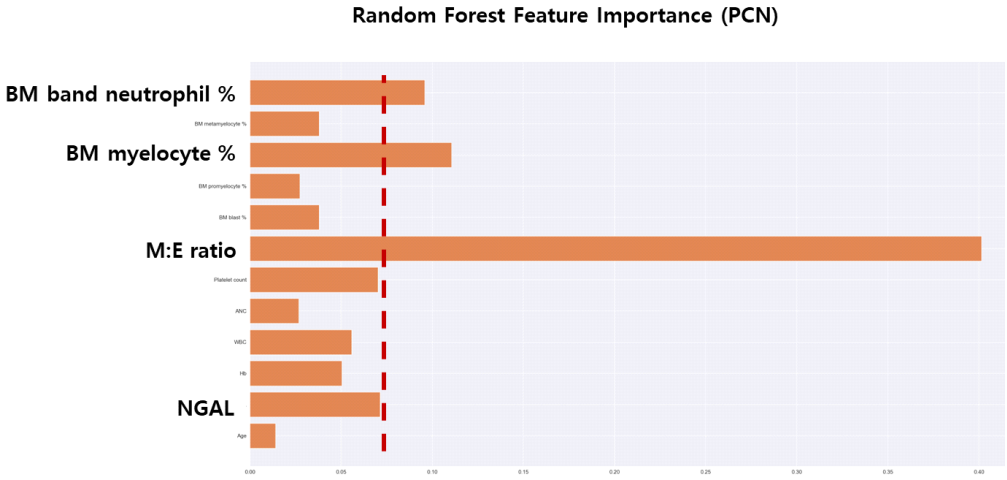

Fig. 12. Feature importance for PCN group's diagnostic prediction by RF.

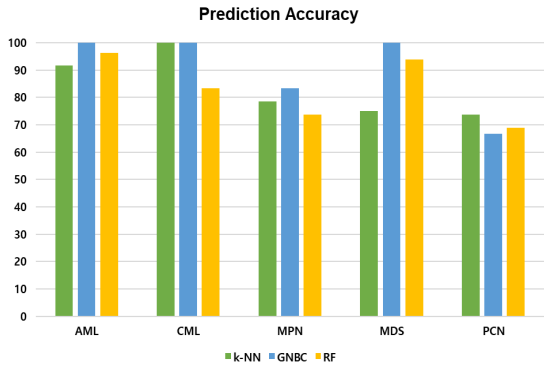

Fig. 13. Accuracy comparison of blood cancer diagnosis predictions without NGAL feature.

4.6 요약

혈액암 진단 예측과 NGAL 지표의 유효성 검증을 기존의 기계학습 기법들을 사용하여 진행하였다. 혈액암의 다섯 가지 종류 중 AML, CML, MDS의 경우 평균 95% 수준의 진단 예측 정확도를 보였으나 MPN과 PCN의 경우 세 가지 분류기 모두 상대적으로 낮은 평균 74% 수준의 진단 예측 정확도를 확인했다. 또한, RF에서 single tree 기준값으로 사용되는 지표 종류 및 특성 중요도 결과를 통해 NGAL 외에 다른 혈액학적 측정 지표들인 M:E ratio, BM myelocyte % 등이 분류에서 더욱 중요도를 가지는 것을 확인했다. 이를 통해 혈액암 진단 예측에서의 NGAL과의 연관성이 높을 것이라는 예상과 달리 절대적인 지표가 아님을 확인하였다. 마지막으로 Fig. 13에서 보여주듯이 세 가지 분류기 중 적은 데이터에도 효율적이라 알려진 GNBC가 일반적으로 성능이 높은 것을 확인할 수 있었다. 따라서, 사용된 데이터에 한정적으로 NGAL의 예측진단과 관련한 유효성 및 보조수단으로써 적합한 분류기가 GNBC임을 확인할 수 있었으며, 추후 추가적인 데이터 활용에 참고할 수 있는 가능성을 제시하였다.

5. 결 론

의료분야에 대한 인공지능 활용은 많은 잠재력을 가지고 있다. 인공지능 기반 질병 진단 예측과 질병 관련 지표들에 대한 유효성 검증을 추가적인 보조수단으로써 활용하여 의료분야에 기여할 수 있다. 본 논문에서는 기존의 기본적인 기계학습 기법들을 분류기로써 사용하여 혈액암 검체 데이터에 대해 진단 예측 성능 비교 및 NGAL 과 혈액암에 대한 연관성을 확인하기 위해 검체 데이터에서의 NGAL 유무에 따른 예측 정확도를 비교하였다. 혈액암의 다섯 가지 종류 중 AML, CML, MDS의 경우 평균 95% 수준의 진단 예측 정확도를 보였으나 MPN과 PCN의 경우 세 가지 분류기 모두 상대적으로 낮은 평균 74% 수준의 진단 예측 정확도를 확인했다. 또한, RF에서 기준값으로 사용되는 지표들과 NGAL의 비교 및 RF 특성 중요도 결과로부터 진단 예측에서의 NGAL과 혈액암에 대한 연관성이 높을 것이라는 예상과 달리 절대적인 지표가 아님을 확인하였다. 마지막으로, 진단 예측에 대해 k-NN, GNBC, RF 분류기 중 GNBC

를 사용했을 때 가장 높은 정확도 결과로부터 GNBC가 적합한 분류기임을 확인하고 사용된 데이터의 분포도 차이가 존재함을 확인할 수 있었다. 제한적인 검체 데이터 개수로 인한 기계학습 정확도 및 신뢰도 확보에 한계가 존재하지만, 간단한 예측 진단 결과로부터 NGAL 지표 유효성 및 분류기 종류 적합성에 대한 확인이 가능하였다. 이를 통해 질병 예측 진단 및 검체 데이터 지표 분석의 보조수단으로써 사용될 수 있는 인공지능 기법들의 활용성과 가능성을 확인할 수 있었다.

REFERENCE

[1] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, and S. Ma, et al., "Artificial Intelligence in Healthcare: Past, Present and Future," *Stroke and Vascular Neurology*, Vol. 2, Issue 4, e000101, 2017.

[2] J. Wiens and E.S. Shenoy, "Machine Learning for Healthcare: on The Verge of a Major Shift in Healthcare Epidemiology," *Clinical Infectious Diseases*, Vol. 66, Issue 1, pp. 149-153, 2018.

[3] T. Davenport and R. Kalakota, "The Potential for Artificial Intelligence in Healthcare," *Future Healthcare Journal*. Vol. 6, No. 2, pp. 94-98, 2019.

[4] H. Long, S.-H. Lee, S.-G. Kwon, and K.-R. Kwon, "A Deep Learning Method for Brain Tumor Classification Based on Image Gradient," *Journal of Korea Multimedia Society*, Vol. 25, No. 8, pp. 1233-1241, 2022.

[5] A. Kosvyra, C. Maramis, and I. Chouvarda, "Developing an Integrated Genomic Profile for Cancer Patients with The Use of NGS Data," *Emerging Science Journal*, Vol. 3, No. 3, pp. 157-167, 2019.

[6] C.H. Cho and J. Cha, "Analysis of Neutrophil Gelatinase-Associated Lipocalin, Vascular Endothelial Growth Factor, and Soluble Receptor for Advanced Glycation End-Products in Bone Marrow Supernatant in Hematologic Malignancies," *Clinical Biochemistry*, Vol. 80, pp. 19-24, 2020.

[7] O. Altay and M Ulas, "Prediction of The Autism Spectrum Disorder Diagnosis with Linear Discriminant Analysis Classifier and K-Nearest Neighbor in Children," *2018 6th International Symposium on Digital Forensic and Security,* pp. 1-4, 2018.

[8] P. Sinha and P. Sinha, "Comparative Study of Chronic Kidney Disease Prediction Using KNN and SVM," *International Journal of Engineering Research and Technology,* Vol. 4, Issue 12, pp. 608-612, 2015.

[9] K. Mittal, G. Aggarwal, and P. Mahajan, "Performance Study of K-Nearest Neighbor Classifier and K-Means Clustering for Predicting The Diagnostic Accuracy," *International Journal of Information Technology,* Vol. 11, Issue 3, pp. 535-540, 2019.

[10] B. Dai, R.C. Chen, S.Z. Zhu, and W.W. Zhang, "Using Random Forest Algorithm for Breast Cancer Diagnosis," *2018 International Symposium on Computer, Consumer and Control*, pp. 449-452, 2018.

[11] P.J. Moore, T.J. Lyons, and J. Gallacher, "Random Forest Prediction of Alzheimer's Disease Using Pairwise Selection from Time Series Data," *Public Library on Science One*, Vol. 14, Issue 2, e0211558, 2019.

[12] D. Yao, Y. Jing, and Z. Xiaojuan, "A Novel Method for Disease Prediction: Hybrid of Random Forest and Multivariate Adaptive Regression Splines," *Journal of Computers,* Vol. 8, Issue 1, pp. 170-177, 2013.

[13] M. Langarizadeh and F. Moghbeli, "Applying Naive Bayesian Networks to Disease Prediction: a Systematic Review," *Acta Informatica Medica,* Vol. 24, Issue 5, pp. 364-369, 2016.

[14] M.A. Jabbar and S. Shirina, "Heart Disease Prediction System Based on Hidden Naïve Bayes Classifier," *2016 International Conference on Circuits, Controls, Communications and Computing,* pp. 1-5, 2016.

[15] D. Dumitru, "Prediction of Recurrent Events in Breast Cancer Using The Naive Bayesian Classification," *Annals of the University of Craiova-Mathematics and Computer Science Series*, Vol. 36, Issue 2, pp. 92-96, 2009.

[16] K. Vembandasamy, R. Sasipriya, and E. Deepa, "Heart Diseases Detection Using Naive Bayes Algorithm," *International Journal of Innovative Science, Engineering & Technology*, Vol. 2, Issue 9, pp. 441-444, 2015.

[17] B.A. Thakkar, I.H. Mosin, and M.A. Desai, "Health Care Decision Support System for Swine Flu Prediction Using Naïve Bayes Classifier," *2010 International Conference on Advances in Recent Technologies in Communication and Computing Institute of Electrical and Electronics Engineers,* pp. 101-105, 2010.

[18] S.D. Jadhav and H.P. Channe, "Comparative Study of K-NN, Naive Bayes and Decision Tree Classification Techniques," *International Journal of Science and Research,* Vol. 5, Issue 1, pp. 1842-1845, 2016.

[19] S.S. Soni, C. Dinna, B. Ilona, Y.C. Chang, N. Federico, and L. Paolo, et al., "NGAL: a Biomarker of Acute Kidney Injury and Other Systemic Conditions," *International Urology and Nephrology*, Vol. 42, Issue 1, pp. 141-150, 2010.

[20] S. Chakraborty, S. Kaur, S. Guha, and S.K. Batra, "The Multifaceted Roles of Neutrophil Gelatinase Associated Lipocalin (NGAL) in Inflammation and Cancer," *Biochimica et Biophysica Acta - Reviews on Cancer,* Vol. 1826, Issue 1, pp. 129-169, 2012.

[21] H. Takizawa and M.G. Manz, "Impact of Inflammation on Early Hematopoiesis and The Microenvironment," *International Journal of Hematology*, Vol. 106, Issue 1, pp. 27-33, 2017.

[22] S.Y. Kristinsson, M. Bjorkholm, M. Hultcrantz, A.R. Derolf, O. Landgren, L. and R. Goldin, "Chronic Immune Stimulation Might Act as a Trigger for The Development of Acute Myeloid Leukemia or Myelodysplastic Syndromes," *Journal of Clinical Oncology*, Vol. 29, Issue 21, pp. 2897-2903, 2011.

[23] S.Y. Kristinsson, O. Landgren, J. Samuelsson, M. Bjorkholm, and L.R. Goldin, "Autoimmunity and The Risk of Myeloproliferative Neoplasms," *Haematologica,* Vol. 95, Issue 7, pp. 1216-1220, 2010.

[24] C.H. Cho, J. Yoon, D.S. Kim, S.J. Kim, H.J. Sung, and S.R. Lee, "Association of Peripheral Blood Neutrophil Gelatinase Associated Lipocalin Levels with Bone Marrow Neutrophil Gelatinase Associated Lipocalin Levels and Neutrophil Count in Hematologic Malignancy," *Journal of Clinical Laboratory Analysis*, Vol. 33, Issue 6, e22920, 2019.

[25] L.M. Zouhal and T. Denoeux, "An Evidence-Theoretic K-NN Rule with Parameter Optimization," *Institute of Electrical and Electronics Engineers Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. 28, Issue 2, pp. 263-271, 1998.

[26] Y. Huang and L. Li, "Naive Bayes Classification Algorithm Based on Small Sample Set," *2011 Institute of Electrical and Electronics Engineers International Conference on Cloud Computing and Intelligence Systems*, pp. 34-39, 2011.

[27] B.W. Silverman and M.C. Jones, "E. Fix and J.L. Hodges (1951): an Important Contribution to Nonparametric Discriminant Analysis and Density Estimation," *International Statistical Review / Revue Internationale de Statistique,* Vol. 57, No. 3, pp. 233-247, 1989.

[28] T. Cover and P. Hart, "Nearest Neighbor Pattern Classification," *Institute of Electrical and Electronics Engineers Transactions on Information Theory*, Vol. 13, No. 1, pp. 21-27, 1967.

[29] H.T. Kam, "Random Decision Forests," *Pro-*

ceedings of 3rd International Conference on Document Analysis and Recognition, Vol. 1, pp. 278-282, 1995.

[30] L. Ceriani and P. Verme, "The Origins of the Gini Index: Extracts from Variabilità e Mutabilità (1912) by Corrado Gini," *The Journal of Economic Inequality*, Vol. 20, No. 3, pp. 421-443, 2012.

[31] G.H. John and P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers," *arXiv Preprint*, arXiv:1302.4964, 2013.

[32] H. Takizawa and M.G. Manz, "Impact of Inflammation on Early Hematopoiesis and the Microenvironment," *Progress in Hematology,* Vol. 106, pp. 27 - 33, 2017.

[33] D.A. Adjeroh, M. Ryynanen, and K.C. Nwosu, "Multimedia Database Management Issues," *Journal of Korea Multimedia Society*, Vol. 4, No. 3, pp. 24-33, 1997.

정 재 승

2015년 서울시립대학교 전자전기컴퓨터공학부 (공학사)
2017년 고려대학교 신재생에너지 (공학석사)
2022년 과학기술연합대학원 나노-정보 융합 (공학박사)

Interest : Neuromorphic engineering, Bayesian neural network

주 현 수

2003년 한국과학기술연구원 원자력및양자공학과 (공학사)
2007년 University of Illinois at Urbana-Champaign (공학석사)
2010년 University of Illinois at Urbana-Champaign (공학박사)

Interest : Neuromorphic engineering, Probabilistic neural network

조 치 현

2003년 고려대학교 의과대학 (학사)
2010년 고려대학교 의과대학 진단검사의학 (석사)
2013년 고려대학교 의과대학 진단검사의학 (박사)

Interest : Statistical diagnostic laboratory medicine, Hematologic malignancy