ISSN: 3022-5388
JKAIA website: https://accesson.kr/jkaia
DOI: http://doi.org/10.24225/jkaia.2024.2.1.7

앙상블 기법을 활용한 RNA-Sequencing 데이터의 폐암 예측 연구

A Study on Predicting Lung Cancer Using RNA-Sequencing Data with Ensemble Learning

Geon AN[1], JooYong PARK[2]

Received: May 20, 2024. Revised: June 14, 2024. Accepted: June 14, 2024

Abstract

In this paper, we explore the application of RNA-sequencing data and ensemble machine learning to predict lung cancer and treatment strategies for lung cancer, a leading cause of cancer mortality worldwide. The research utilizes Random Forest, XGBoost, and LightGBM models to analyze gene expression profiles from extensive datasets, aiming to enhance predictive accuracy for lung cancer prognosis. The methodology focuses on preprocessing RNA-seq data to standardize expression levels across samples and applying ensemble algorithms to maximize prediction stability and reduce model overfitting. Key findings indicate that ensemble models, especially XGBoost, substantially outperform traditional predictive models. Significant genetic markers such as ADGRF5 is identified as crucial for predicting lung cancer outcomes. In conclusion, ensemble learning using RNA-seq data proves highly effective in predicting lung cancer, suggesting a potential shift towards more precise and personalized treatment approaches. The results advocate for further integration of molecular and clinical data to refine diagnostic models and improve clinical outcomes, underscoring the critical role of advanced molecular diagnostics in enhancing patient survival rates and quality of life. This study lays the groundwork for future research in the application of RNA-sequencing data and ensemble machine learning techniques in clinical settings.

Keywords : Lung Cancer, RNA-sequencing, Gene Expression , Ensemble Learning, Machine Learning

Major Classification Code : Artificial Intelligence, etc

1. Introduction

폐암은 전 세계적으로 가장 일반적이며 치명적인 암이다. 세계보건기구에 따르면, 폐암은 전 세계 암 사망자 중 가장 큰 비율을 차지하며, 매년 약 180만 명이 폐암으로 사망하고 있다 (World Health Organization, 2023). 초기의 폐암은 대부분 증상이 없거나, 증상이 나타나더라도 감기나 독감과 같은 흔한 질환의 증상과 유사하여 쉽게 관과 할 수 있다. 또한, 폐암 진단을 위한 X-ray 방식은 작은 병변이나 폐의 이상을 정확히 확인하는데 한계가 있기에 폐암이 대부분 매우 늦은 단계에서 발견된다 (Li et al., 2020).

1 First Author. Undergraduate Student, Department of Medical IT, Eulji University, Republic of Korea, Email: geon0078@g.eulji.ac.kr
2 Corresponding Author. Assistant Professor, Department of Big Data Medical Convergence, Eulji University, Republic of Korea, Email: jy.park@eulji.ac.kr

폐암의 예측은 생존율을 크게 향상시킬 수 있다. 초기에 폐암 치료를 시작하면 더 효과적이며, 완치 가능성이 높아진다 (Zappa & Mousa, 2016). 최근 연구와 기술의 발달은 폐암의 예측을 위한 새로운 방법을 제시하고 있으며, 적절한 검사를 통해 고위험군에 대한 발병률을 예측하여 이른 시기에 완치를 할 수 있는 가능성을 열어주고 있다 (Nooreldeen & Bach, 2021).

RNA sequencing(RNA-seq)은 전사체(transcriptome)의 양적 및 질적 분석을 가능하게 하는 기술이다. 전사체는 특정 시점에서 세포, 조직 또는 개체에서 발현되는 모든 Ribonucleic Acid (RNA)분자를 포함하며 RNA-seq는 유전자 발현의 차이를 정량화하고, 새로운 전사 단위를 식별하며, alternative splicing 및 변이를 감지하는 데 사용된다. RNA는 DNA에서 전사되며, 이 RNA는 Messenger RNA(mRNA), Ribosomal RNA (rRNA), Transfer RNA(tRNA) 등 다양한 유형의 RNA로 분류된다. 그 중 mRNA는 세포 내에서 유전자에서 전사된 후에 단백질 합성에 직접 참여하는 RNA이기 때문에, RNA-seq data 분석을 위해 주로 mRNA를 대상으로 한다. mRNA를 추출하기 위해 RNA로부터 rRNA와 tRNA를 분리하는 일반적인 방법은 Poly-A tail 선택이다. 대부분의 진학 생물의 mRNA는 3'-end에 Poly-A tail를 가지고 있는 반면, 다른 RNA 유형들은 이런 특징을 갖지 않는다. 이를 이용하여 Poly-T 염기 서열이 부착된 칼럼이나 비드를 사용하여 mRNA만을 선택적으로 붙잡을 수 있다. Poly-A tail 있는 mRNA가 Poly-T 서열에 결합하여 다른 RNA 유형들은 제거된다. 그런 다음, mRNA는 complementary DNA(cDNA)로 역전사되고, 이 cDNA는 시퀀싱 라이브러리를 구성하는 데 사용된다. 이 라이브러리는 시퀀싱을 통해 읽히며 각 읽기(read)는 기존의 mRNA 분자를 나타내고 이를 통해 유전자 발현을 확인할 수 있다. 이렇게 생성된 데이터는 컴퓨터로 전송되어 분석되며 분석된 데이터는 유전자 발현 수준을 정량화하고, 새로운 전사 단위를 식별하며, alternative splicing events를 감지하는 데 사용된다. 이렇게 얻은 정보는 질병 예측, 진단, 치료 등에 활용될 수 있으며 최근 몇 년 동안 생물학적 연구와 질병 진단에 혁신을 가져왔다. 특히, 암 연구에서 강력한 도구로 자리 잡아왔다 (Ergin et al., 2022).

앙상블 학습 기법은 머신 러닝의 한 분야로, 여러 개별 모델의 예측을 결합하여 단일 모델의 성능을 초과하는 예측 성능을 달성하는 방법이다. 이 방식은 다양한 예측 모델의 강점을 통합하고, 모델들의 약점을 상쇄함으로써, 전체적으로 더 견고하고 신뢰할 수 있는 예측 결과를 제공한다. 앙상블 학습은 분류, 회귀, 그리고 클러스터링 문제에 널리 적용되며, 특히 복잡하고 불확실성이 높은 의료 데이터 분석에 매우 효과적이다. 앙상블 기법은, 주로 배깅(Bagging)과 부스팅(Boosting)의 두 가지 주요 유형으로 나뉜다. 먼저 배깅은 원본 데이터 세트에서 여러 개의 랜덤하게 선택된 서브 세트 샘플을 생성하여 각 서브 세트에 독립적인 모델을 학습시킨다. 이렇게 학습된 각각의 모델은 동일한 확률로 결정에 기여하며, 최종 예측은 이 모델들의 결과를 평균화하거나 투표(다수결)를 통해 이루어진다. 배깅의 주된 장점은 모델의 분산을 감소시켜 과적합을 줄이는 데 효과적이라는 점이다. 부스팅은 약한 예측 모델들을 순차적으로 학습시키는 기법으로, 각 단계에서 이전 모델들의 예측이 틀린 데이터에 더 큰 가중치를 부여하며, 모델을 학습시킨다. 이 과정을 통해, 이전 모델의 오류를 차례대로 수정해 나가면서 점진적으로 예측력을 높여간다.

폐암 예측 연구에서 앙상블 기법의 활용은 상당한 주목을 받아왔다. 특히, RNA-seq data와 같은 고차원의 생물학적 데이터를 처리할 때, 앙상블 기법은 모델의 성능과 해석 가능성을 크게 향상시키는 데 기여하였다. 여러 연구에서는 랜덤 포레스트와 그래디언트 부스팅 모델을 폐암 환자의 유전자 발현 데이터에 적용하여, 질병의 진단, 예후 예측, 그리고 치료 반응 예측에 시도하였다 (Baradaran Rezaei et al., 2023). 이번 연구는 RNA-seq data를 앙상블 모델에 적용해서 폐암 발생에 어떤 유전체 발현이 영향을 미치는지 탐색하고 예측 인자를 발굴하고자 한다.

2. Related Research

관련 연구로 Park et al. (2021)은 RNA 시퀀싱 데이터를 사용하여 Ulcerative Colitis(UC)와 Crohn' s Disease(CD)을 구별할 수 있는 기계학습 모델을 개발하였다. 연구팀은 염증성 장 질환 환자의 내시경 생검 조직에서 RNA-seq data를 수집하고, 인간 참조 게놈(GRCh38)에 매핑하여 19,596개의 단백질 코딩 유전자를 포함하는 유전자 모델을 정량화했다. 비지도 학습 모델은 CD와 UC 두 가지 클래스를 명확하게 구분하는 모습을 보여주었으며 지도 학습 모델에서는 Partial Least Squares Discriminant Analysis(PLS-DA)을 사용하여 영향력이 큰 특성들을 데이터 세트에서 배제하여 염증성 CD와 염증성 UC를 구별할 수 있었다. 결과적으로 전체 오류율은 0.147로 매우 낮았으며, 이 연구는 RNA-seq data 분석과 Machine Learning(ML) 기법을 활용하여 UC와 CD를 효과적

으로 구별할 수 있는 방법론을 제시하였다.

Bostanci et al. (2023)는 ML과 Deep Learning(DL) 모델을 이용하여 대장암의 진단과 예후를 예측하는 방법을 탐구하였다. 사용된 데이터는 건강한 사람과 대장암 환자의 세포 외 소포체의 RNA-seq data를 분석한 것으로, 다양한 ML 및 DL 분류기를 활용하여 대장암 단계와 암 존재 여부를 예측하였다. 기존의 ML 분류기로는 K-Nearest Neighbors(KNN), Logistic Model Trees(LMT), Random Trees(RT), Random Committees(RC), Random Forests(RF) 등이 사용되었으며, 더불어 1-D Convolutional Neural Networks(1-D CNN), Long Short-Term Memory networks (LSTM), Bidirectional LSTMs(BiLSTM)과 같은 DL 모델들도 평가하였다. 실험 결과, 대장암 예측에서는 RC, LMT, RF 모델이 97.33%의 가장 높은 정확도를 보였으며, 대장암 단계 분류에서는 RF의 정확도가 97.33%로 가장 높은 성능을 보였다. 1-D CNN은 대장암 예측에서 97.67%로 가장 높은 정확도를 나타내었고, BiLSTM은 대장암 단계 분류에서 98%의 정확도로 가장 좋은 결과를 보였다. 종합적으로, 이 연구는 RNA-seq data를 활용하여 대장암의 진단과 예후 예측을 위한 다양한 기계학습과 심층학습 방법을 적용하여 높은 정확도를 달성할 수 있음을 보여주었으며 이러한 접근 방식은 임상 연구에 있어서 유용한 도구가 될 수 있고 향후 더 많은 생물정보학적 데이터와의 통합을 통해 질병의 조기 진단과 치료에 기여할 수 있을 것이라고 주장하였다.

Piao et al. (2014)는 앙상블 분류 알고리즘을 사용하여 전립선암을 정확하게 예측하고자 하는 연구를 진행하였다. 특히, Support Vector Machine(SVM)을 기반 분류기로 사용하는 앙상블 방법을 채택하여 높은 차원의 RNA-seq data를 전립선암을 진단하는 모델을 개발하였다. 이 연구에서는 다양한 기계 학습 기법과 유전자 선택 방법을 통합하여 정확도를 높이는 방법을 제안하고 있다. 특히, 특징 선택 알고리즘과 다양한 분류기를 결합한 앙상블 방법이 중요한 역할을 하며, 이를 통해 더 나은 학습 성능을 달성할 수 있다 주장한다. 이러한 방법들을 실제 데이터 세트와 시뮬레이션 된 데이터 세트에 적용하여 모델의 성능을 평가하였고, 이러한 접근 방식이 전립선암 진단의 정확도를 개선할 수 있음을 보여주고 있다. 또한, 다양한 정규화 방법과 RNA-seq data를 이용한 시뮬레이션 기법을 통해 데이터의 처리 과정에서 발생할 수 있는 편향을 최소화하고자 하였으며, 고차원의 유전자 발현 데이터로부터 후보 유전자를 식별함으로써 전립선암의 이해를 높이고 이를 기반으로 한 질병 예측에 기여하고자 하였다.

3. Research

3.1. Collecting RNA-seq Data

본 연구는 National Cancer Institute(NCI)와 National Human Genome Research Institute Home(NHGRI)에서 주도하는 The Cancer Genome Atlas-Lung Squamous Cell Carcinoma(TCGA-LUSC)프로젝트에서 제공하는 데이터를 기반으로, 총 504명의 데이터 및 56,907개의 RNA-seq data를 포함하고 있다. 이 데이터는 폐의 편평상피 세포암 환자와 정상인 들로부터 수집된 대규모 데이터를 담고 있으며, 폐암의 발생과 진행에 중요한 유전자와 경로를 식별하는 데 중요한 기여한다. 앙상블 모델을 처리하기 전 예측 정확도를 최대화 위해 다음과 같은 전처리 과정을 수행하였다.

3.2. Preprocessing RNA-seq Data

첫 단계로, 모든 유전자 발현 데이터는 Transcripts Per Million (TPM) 방식으로 정규화되어, 다양한 샘플 간 발현 수준을 일관되게 비교할 수 있도록 했다. TPM은 RNA-seq data에서 유전자 발현 수준을 정량화 하는 방법 중 하나로 유전자의 길이와 시퀀싱 깊이를 보정하여, 서로 다른 샘플 간의 유전자 발현을 비교 가능하게 해준다.

TPM을 구하기 위해 먼저 RPK를 구하는 과정이 수행된다.

$$RPK = \frac{Read\ Count}{Gene\ Length\ in\ kilobases} \quad (1)$$

RNA-seq에서 얻은 각 유전자의 read counts는 유전자의 길이에 영향을 받는다. 유전자 길이가 길면 그 유전자에 매핑 될 수 있는 리드의 수가 많아지므로 수식 (1)을 통해 유전자 길이를 보정한다. RPK 정규화 값은 각 샘플의 시퀀싱 깊이와 유전자 길이에 따라 조정되지만, 샘플 간 비교를 위해서는 총 시퀀싱 출력을 표준화하는 추가 단계가 필요하다.

$$TPM = \frac{(RPK)}{\sum RPK\ of\ all\ genes} \times 10^6 \quad (2)$$

그 다음 TPM 은 각 유전자의 발현이 전체 유전자 발현 중 어느 정도의 비율을 차지하는지를 보여준다. TPM

정규화는 수식 (2)와 같은 계산을 통해 각 샘플의 발현 데이터를 표준화하고, 생물학적 또는 실험적 변이에 의한 영향을 최소화하여 데이터의 비교 분석을 보다 정확하게 수행할 수 있도록 돕는다.

3.2.1. Logistic Regression Analysis

Logistic regression analysis 는 예측 모델링 및 데이터 분석에 사용되는 통계 기법이다. 이 분석은 종속 변수가 이진일 때 사용되며, 독립 변수가 특정 범주형 결과에 미치는 영향을 추정할 수 있으며, 이 분석 방법을 통해 p-value 를 계산하였다.

3.2.2. Bonferroni Correction

Bonferroni correction 은 다중 검정 문제에서 유의수준을 보정하는 방법이다. 일반적으로 단일 가설 검정에서 유의수준 α 을 0.05 또는 0.01 로 설정한다. 하지만 다중검정에서는 여러 개의 가설을 동시에 검정하기 때문에 유의수준을 높게 설정하면 잘못된 결론을 내릴 가능성이 있기에 이를 방지하기 위한 Bonferroni correction 은 다중 검정 문제에서 각 가설에 대한 유의수준을 조정하여 전체적인 오류를 제어한다. Bonferroni correction 을 사용하여 보정된 유의수준을 구하는 방법은 다음과 같다.

$$\alpha_{\text{Bonferroni}} = \frac{\alpha}{m} \quad (1)$$

수식 (1)에서 α는 0.05 나 0.01 로 설정되며, m은가설의 개수를 나타낸다. 이렇게 보정된 유의수준을 사용하면 다중 검정에서 각 가설에 대한 보정된 유의수준을 적용하여 전체적인 유의성을 보다 엄격하게 제어할 수 있다.

3.2.3. Fold Change

유전자 발현 분석에서 Fold Change 계산은 특정 조건 또는 처리에 대한 유전자의 발현 변화를 정량적으로 평가하는 데 매우 중요한 지표이다. Fold Change 는 한 상태에서의 유전자 발현 수준과 다른 상태에서의 발현 수준을 비교하여, 발현이 얼마나 증가하거나 감소했는지를 나타낸다. Fold Change 는 다음과 같은 단계로 계산된다. 먼저 각 조건에 대해 유전자의 평균 발현 수준을 계산한다. 이때, TPM 을 통해 얻은 값들을 사용하여 각 샘플의 발현 수준을 평균화한다. 다음 실험군의 평균 발현 수준을 대조군의 평균 발현 수준으로 나누어 Fold Change 를 계산한다. 발현 수준이 증가한 경우, Log_2(Fold Change) 값은 1 이상이 될 것이며, 발현 수준이 감소한 경우에는 1 미만이 된다.

3.3. Extracting Training Data

유전자 발현 연구에서 유의미한 차이를 보이는 유전자를 식별하는 것은 매우 중요한 단계이다. 본 연구에서는 통계적 유의성과 생물학적 의미를 모두 고려하여, 폐의 편평상피세포암과 관련된 유전자를 특정하기 위해 Bonferroni correction 을 수행한 p-value 가 0.01 이하이며, $\text{Log}_2(\text{Fold Change})$ 가 2 이상이거나 -2 이하인 경우를 만족하는 유전자 2,774 개의 대해 유의미하다고 판단하고, 학습데이터로 추출하였다.

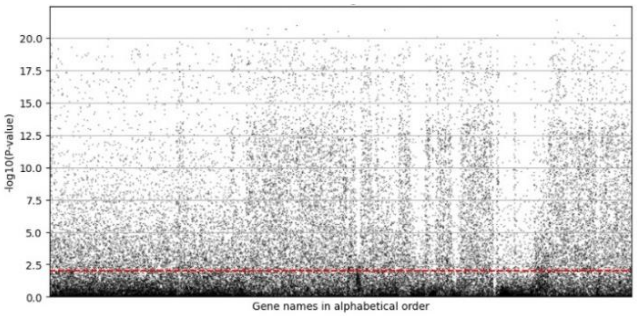

Figure 1: Scatter Plot of -log10(P-value)

3.4. Ensemble Artificial Intelligence Model

3.4.1. Random Forest

Random Forest(RF)는 다양한 결정 트리를 조합하여 사용하는 앙상블 머신 러닝 기법으로, 분류 및 회귀 문제에 효과적으로 사용된다. 이 알고리즘은 각각의 결정 트리가 데이터 세트의 서브 세트에서 독립적으로 학습하고, 그 결과를 통합하여 최종 예측을 도출함으로써 단일 결정 트리의 성능보다 더 높은 정확도와 안정성을 제공한다. 결정 트리의 각 노드는 데이터를 분할하는 결정 포인트로서 기능하며, 노드에서는 하나의 특성을 기준으로 데이터를 두 그룹으로 나눈다. 이 분할은 데이터의 특성에 따라 최적의 질문을 선택하고, 각 노드에서 최적의 분할을 찾는 과정을 통해 이루어진다. RF 의 핵심 메커니즘은 Bootstrap Sampling(BS) Random Selection of Features(RSF)에 기반을 두고 있다. 이 두 기법을 통해 원본 데이터에서 복원 추출 방식으로 각 트리의 학습 데이터를 생성하며, 트리가 성장하는 과정에서는 전체 특성 집합이 아닌 무작위로 선택된 소수 특성을 사용하여 분할을 결정한다. 이러한 방법은 각 트리가 서로 다른 특성의 조합을 기반으로

학습하도록 하여 트리 간의 상관관계를 감소시키고, 전체 모델의 과적합 위험을 줄이는 데 기여한다.

3.4.2. eXtreme Gradient Boosting

eXtreme Gradient Boosting(XGB)는 기계 학습에서 널리 사용되는 고급 앙상블 기법 중 하나이다. XGB 는 부스팅 방식을 활용하여 여러 약한 학습기, 주로 결정 트리를 순차적으로 학습시키면서 각 단계에서 이전 모델의 오류를 수정해 나가는 구조로 이루어져 있다. XGB 는 트리를 순차적으로 구축하며 이전 트리의 오류를 계속해서 보정해 나가며 이 점에서 RF 와 차이점이 있다. 이 방식은 모델의 정확도를 높이는 데 기여하며, 특히 노이즈가 많은 데이터에서 강력한 성능을 발휘할 수 있도록 한다. XGB 의 또 다른 특징은 정규화 매개변수를 포함하여 과적합을 방지하는 것이며, 모델의 일반화 성능을 향상시키는 것이다. 또한, 병렬 처리 및 하드웨어 최적화를 지원하여 대규모 데이터 세트를 빠르게 처리할 수 있는 장점이 있다.

3.4.3. Light Gradient Boosting Machine

Light Gradient Boosting Machine(Light GBM)은 Microsoft 에 의해 개발된 고급 Gradient Boosting 프레임워크이다. 이 기법은 특히 대규모 데이터 세트에 대한 처리 속도와 효율성 측면에서 매우 우수하며, 머신 러닝에서 분류와 회귀 문제를 해결하는 데 널리 사용된다. LightGBM 은 Gradient-Based One-Side Sampling(GOSS) 과 Exclusive Feature Bundling(EFB)이라는 두 가지 주요 기술을 통해 데이터 처리의 속도와 효율성을 극대화한다. GOSS 은 기계 학습에서 모델 학습 속도를 높이기 위한 한 가지 기법으로 모델이 특정 데이터 포인트에 대해 예측하는 방향으로만 샘플을 추출하여 학습 속도를 높이는 방식이다. 기존의 랜덤 샘플링은 매 학습 반복마다 모든 데이터 포인트에서 랜덤하게 샘플을 추출하여 학습 속도를 느리게 만드는 반면, GOSS 에서는 모델이 현재 가장 잘못 예측하는 데이터 포인트 즉 약점에 초점을 맞추어 샘플을 추출한다. 이 방법을 사용하면 모델의 학습이 진행됨에 따라 모델이 여전히 잘못 예측하는 데이터에 대해 더 자주 학습하게 모델의 학습 속도를 높이고 수렴 속도를 개선하는 데 도움을 준다. EFB 은 Feature Selection 의 기법 중 하나로, 모델의 복잡성을 줄이고 성능을 개선하기 위해 사용된다. 이 기법은 데이터의 특성을 묶어서 하나의 새로운 특성으로 결합하는 방식으로 작동한다. 일반적으로, 데이터 세트에는 다양한 특성이 존재하는데, 이 중에서 일부 특성은 서로 상관관계가 높아서 중복되거나 유사한 정보를 제공한다. 이러한 중복된 특성들은 모델을 더 복잡하게 하고, 과적합을 유발할 수 있기 때문에 EFB 는 이러한 중복된 특성을 하나의 묶음으로 만들어서 모델이 이를 하나의 특성으로 취급하도록 한다. 이렇게 함으로써, 모델은 데이터의 정보를 더 효율적으로 활용하면서도 복잡성을 줄일 수 있다. 또한, 묶음으로 만들어진 새로운 특성은 데이터의 중요한 측면을 잡아내는 데 도움을 준다.

3.5. Hyperparameter Optimization

모델의 성능을 극대화하고 일반화 능력을 항상 시키기 위해 다음과 같은 하이퍼파라미터 최적화 과정을 수행하였다. 먼저 K-Folds Cross Validation 은 모델의 성능을 평가하고 일반화 능력을 검증하기 위한 통계적 방법이다. 이 방법은 전체 데이터 세트를 K 개의 동일한 크기의 Fold 로 나누고, 이 중 하나의 Fold 를 테스트 데이터로, 나머지 K-1 개의 Fold 를 학습 데이터로 사용하여 모델을 훈련시킨다. 이 과정을 Fold 마다 한 번씩, 총 K 번 반복하면서 모델을 테스트하며 각 반복에서 얻은 성능 지표를 평균내어 모델의 최종 성능을 추정한다. 이 방식은 데이터의 모든 샘플이 학습과 검증에 골고루 사용되도록 하여, 데이터의 낭비를 줄이고 모델 평가의 신뢰성을 높인다. K-Folds Cross Validation 이후 Random Search 과정을 수행하였다. Random Search 는 하이퍼파라미터 최적화를 위한 방법 중 하나로 가능한 하이퍼파라미터의 공간 내에서 무작위로 선택한 하이퍼파라미터의 조합을 사용하여 모델을 훈련시키고, 그 성능을 평가한다. 랜덤 서치는 특히 하이퍼파라미터의 차원이 높고, 최적의 조합이 예측하기 어려울 때 유용하게 사용된다.

Table 1: Optimized hyperparameters for each algorithm

Algorithm	Hyper parameter
Random Forest	- n_estimators: 100 - min_samples_split: 10 - min_samples_leaf: 4 - max_samples: 0.7 - max_leaf_nodes: None - max_features: 'auto' - max_depth: 10 - bootstrap: True
XGBoost	- subsample: 0.6 - n_estimators: 100 - min_child_weight: 3 - max_depth: 5 - learning_rate: 0.05

	- gamma: 0 - colsample_bytree: 1.0
LightGBM	- subsample: 0.7 - num_leaves: 20 - min_child_samples: 30 - max_depth: 10 - learning_rate: 0.1 - colsample_bytre: 1.0

Table 1 은 앞서 언급된 두 방법을 활용한 모델의 주요 하이퍼파라미터 설정이다.

3.6. Evaluating Ensemble Models

본 연구에서는 폐암 예측에 사용된 세 가지 기계학습 모델 RF, XGB, 그리고 LightGBM 의 성능을 평가하기 위하여 세 가지 주요 평가 지표를 적용하였다. Mean Squared Error(MSE)은 예측 값과 실제 값의 차이를 제곱한 값의 평균으로 계산된다. 이 지표는 예측 오차의 크기를 측정하며, 값이 낮을수록 예측 정확도가 높음을 의미한다. 본 연구 결과에서 XGB 모델이 0.003 의 MSE 값을 보여 가장 낮은 예측 오차를 나타냈다. Mean Absolute Error(MAE)은 예측 값과 실제 값의 절대값 차이의 평균을 나타낸다. 모델의 예측 값이 실제 값에 얼마나 근접하는지를 직관적으로 보여주며, 마찬가지로 값이 낮을수록 더 좋은 예측 성능을 의미한다. 연구 결과에 따르면 XGB 모델이 0.001 의 MAE 로 가장 낮은 값을 나타냈으며, 훈련 데이터에 과적합 되지 않고 새로운 유전 데이터에 대해서도 일반화되어 적용 가능하다고 판단된다. 결정 계수(R^2)는 모델이 데이터의 분산을 얼마나 잘 설명하는지를 나타내는 지표로, 값이 1 에 가까울수록 모델의 설명력이 높다고 평가된다. 본 연구에서 XGB 는 0.959 의 R^2 값을 보여, 세 모델 중 가장 높은 데이터 설명력을 갖는 것으로 나타났다.

3.7. Key Predictive Genetic Factors

폐암 예측에 가장 큰 영향을 끼치는 유전자를 확인하게 위해 Local Interpretable Mode-agnostic Explanations(LIME)를 활용하였다. 이 알고리즘은 설명하고자 하는 데이터 포인트에 대해 원래 모델의 예측을 구한다음 목표 데이터 포인트와 유사한 샘플 데이터를 생성한다. 이 샘플 데이터는 목표 데이터 포인트의 특성을 약간 변형하여 만들어지며, 생성된 샘플 데이터에 대해 원래 모델의 예측을 수행하고, 목표 데이터 포인트와의 유사도를 계산하여 각 샘플 데이터에 가중치를 부여한다. 이 가중치를 부여한 샘플 데이터와 예측 결과를 사용하여 선형 모델을 학습하며, 이 선형 모델은 목표 데이터 포인트 주변의 데이터에 대한 단순한 근사 모델로, 원래 모델의 복잡한 예측을 이해하기 쉽게 만들어 준다. 마지막으로, 학습된 선형 모델의 특성 중요도를 분석하여 원래 모델의 예측을 설명한다. 각 특성의 중요도는 해당 특성이 예측 결과에 얼마나 큰 영향을 미쳤는지를 보여주며, 이 과정을 통해 LIME은 복잡한 ML 모델의 예측을 단순하고 이해하기 쉽게 설명해 준다.

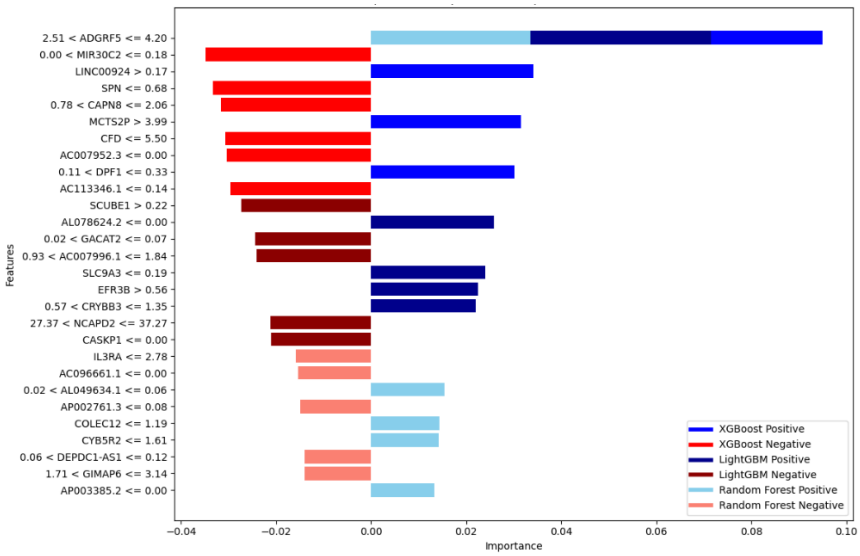

Figure 2: Gene importance using LIME

다음 Figure 2 는 LIME 알고리즘을 통해 RF, XGB, Light BGM 에서 중요 특징으로 추출된 특성을 나타낸다. 먼저 RF 의 특성 중요도를 살펴보면, ADGRF5 가 가장 높은 중요도를 보이며 특히 ADGRF5 값이 2.51 에서 4.20 사이에 있을 때 모델의 예측에 가장 큰 영향을 주었다. 그 다음으로 IL3RA 와 AC096661.1 의 값이 일정 범위 내에 있을 때 모델의 예측에 부정적인 영향을 주었으며 반면 AL049634.1 과 AP002761.3 은 특정 범위 내에 있을 때 모델의 예측에 긍정적인 영향을 주었다. XGB 모델에서 또한 ADGRF5 특성이 가장 높은 중요도를 보였으며 MIR30C2 와 SPN 은 특정 범위 내에 있을 때 모델의 예측에 부정적인 영향을 끼친 반면 LINC00924 와 MCTS2P 는 특정 범위 내에 있을 때 모델의 예측에 긍정적인 영향을 주었다. LightGBM 모델 에서는 SCUBE1 과 AL078624.2 는 특정 범위 내에 있을 때 모델의 예측에 부정적인 영향을 미치는 것으로 나타났으며, 반면 SLC9A3 과 EFR3B 는 특정 범위 내에 있을 때 모델의 예측에 긍정적인 영향을 주었다.

4. Conclusion

이 연구는 폐암 예측을 위해 RNA-seq data 를 활용하여 여러 앙상블 기법들을 적용하였고, 각각의 방법마다 중요도가 높은 유전자 발현은 어떤 것들이 있는지 확인하였으며 각 모델의 성능을 평가하였다. 또한, RF, XGB, 그리고 LightGBM 세 가지 모델을 적용하고, 이들의 예측 성능을 비교 분석하였으며. 세 가지 앙상블 모델을 종합하여 고려할 때, XGB 앙상블 모델이 가장 뛰어났으며, 유전 인자에서는 ADGRF5 유전자가 모든 모델에서 중요한 예측 인자인 것으로 나타났다. 특히 세 모델 모두 ADGRF5 의 값이 2.51 에서 4.20 사이에 있을 때 폐암 예측에 가장 큰 영향을 주었으며 이러한 일관된 결과는 ADGRF5 가 폐암 예측에 있어서 핵심적인 역할을 하는 것으로 보인다. 이전 연구에서 Brown, Filuta, Ludwig, Seuwen, & Jaros (2017)는 ADGRF5 유전자가 폐혈관의 기능 조절에 중요한 역할을 하며, 변이가 있을 경우 Pulmonary Hypertension(PH)가 발생할 수 있다고 주장하였고 Roderburg et al. (2022)는 PH 가 Respiratory Cancer 과 가장 강한 연관성을 갖는다는 것을 발견하였다.

이 연구는 폐암 예측을 위한 RNA-seq data 활용 가능성과 앙상블 기법의 유효성을 보여주고 있으며, 다음 단계로 여러 분야에서의 추가 연구를 제안한다. 먼저 유전자 마커와 다른 생물학적 데이터를 통합하여 진단 모델의 정확성을 향상시켜야 하며 불균형 데이터 세트를 처리하고 모델의 일반화 능력을 높이는 알고리즘을 개발해야 한다. 또한 예측 모델의 임상적 적용 가능성을 다기관 연구를 통해 검증해야 하며 향후 연구에서는 이 결과를 바탕으로 폐암예측 및 치료 개발 전략을 추가로 모색하여 폐암 치료의 진전을 촉진하고 환자의 생존율 및 삶의 질 향상에 기여할 것으로 기대한다.

References

Ali, J., Khan, R., & Ahmad, N. (2012). Random forests and decision trees. *International Journal of Computer Science Issues, 9*(5). Retrieved from https://www.uetpeshawar.edu.pk/TRP-G/Dr.Nasir-Ahmad-TRP/Journals/2012/Random%20Forests%20and%20Decision%20Trees.pdf

Baradaran Rezaei, H., Amjadian, A., Sebt, M. V., et al. (2023). An ensemble method of the machine learning to prognosticate the gastric cancer. *Annals of Operations Research, 328*, 151–192. https://doi.org/10.1007/s10479-022-04964-1

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research, 13*(Feb), 281-305. Retrieved from https://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf

BigOmics Analytics. (2023, March 16). What is TPM? Understanding normalization methods for gene expression. *BigOmics Analytics*. Retrieved from https://bigomics.ch/blog/why-how-normalize-rna-seq-data/

Boateng, E., & Abaye, D. (2019). A review of the logistic regression model with emphasis on medical research. *Journal of Data Analysis and Information Processing, 7*, 190-207. doi: 10.4236/jdaip.2019.74012.

Bostanci, E., Kocak, E., Unal, M., Guzel, M. S., Acici, K., & Asuroglu, T. (2023). Machine learning analysis of RNA-seq data for diagnostic and prognostic prediction of colon cancer. *Sensors, 23*(6), 3080. https://doi.org/10.3390/s23063080

Brown, K., Filuta, A., Ludwig, M. G., Seuwen, K., & Jaros, J. (2017). Epithelial Gpr116 regulates pulmonary alveolar homeostasis via Gq/11 signaling. *JCI Insight, 2*(11), e89704. https://doi.org/10.1172/jci.insight.89704

Czepiel, S. A. (2002). Maximum likelihood estimation of logistic regression models: Theory and implementation. Available at czep.net/stat/mlelr.pdf

Ergin, S., Kherad, N., & Alagoz, M. (2022). RNA sequencing and its applications in cancer and rare diseases. *Molecular Biology Reports, 49*, 2325–2333. https://doi.org/10.1007/s11033-021-06963-0

Gad, A. A., & Balenga, N. (2020). The emerging role of adhesion GPCRs in cancer. *ACS Pharmacology & Translational Science*. https://doi.org/10.1021/acsptsci.9b00093

Gohiya, H., Lohiya, H., & Patidar, K. (2018). A survey of XGBoost system. *International Journal of Advanced Technology and Engineering Research, 8*(7). Retrieved from http://www.ijater.com/Files/aa09b180-add4-4a6d-b234-bc122eb305d4_IJATER_39_07.pdf

Handoyo, S., Pradianti, N., Nugroho, W. H., & Akri, Y. J. (2022). A heuristic feature selection in logistic regression modeling with newton raphson and gradient descent algorithm. *International Journal of Advanced Computer Science and Applications, 13*(3).

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems, 30*. Retrieved from https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html

Li, K., Chen, Y., Sun, R., Yu, B., Li, G., & Jiang, X. (2020). Exploring potential of different X-ray imaging methods for early-stage lung cancer detection. *Journal of Medical Imaging and Radiation Sciences, 5*(2), 173-183. https://dx.doi.org/10.1007/s41605-020-00173-1

Li, W., Yin, Y., Quan, X., & Zhang, H. (2019). Gene expression value prediction based on XGBoost algorithm. *Frontiers in Genetics.* Retrieved from https://www.frontiersin.org/articles/10.3389/fgene.2019.01077/full

Louppe, G. (2014). Understanding random forests: From theory to practice. *Cornell University Library*. Retrieved from https://www.researchgate.net/profile/Gilles-Louppe/publication/264312332_Understanding-Random-Forests_From-Theory-to-Practice/links/54ae38ea0cf2213c5fe427b7/Understanding-Random-Forests-From-Theory-to-Practice.pdf

Midthun, D. E. (2016). Early detection of lung cancer. *F1000Research, 5*, F1000 Faculty Rev-739. https://doi.org/10.12688/f1000research.7313.1

Napierala, M. A. (2012). What is the Bonferroni correction?, *AAOS Now, 40*. Retrieved from https://link.gale.com/apps/doc/A288979427/HRCA?u=anon~94f28a3d&sid=googleScholar&xid=d9841e38

Nooreldeen, R., & Bach, H. (2021). Current and future development in lung cancer diagnosis. *International Journal of Molecular Sciences, 22*(16), 8661. https://doi.org/10.3390/ijms22168661

Park, S.-K., Kim, S., Lee, G.-Y., Kim, S.-Y., Kim, W., Lee, C.-W., Park, J.-L., Choi, C.-H., Kang, S-B., & Kim, T.-O., et al. (2021). Development of a machine learning model to distinguish between ulcerative colitis and Crohn's disease using RNA sequencing data. *Diagnostics, 11*(12), 2365. https://doi.org/10.3390/diagnostics11122365

Piao, Y., Choi, N. H., Li, M., Piao, M., & Ryu, K. H. (2014). Ensemble method for prediction of prostate cancer from RNA-Seq data. *Science Technology*, 51-56.

Roderburg, C., Loosen, S. H., & Hippe, H. J. (2022). Pulmonary hypertension is associated with an increased incidence of cancer diagnoses. *Pulmonary Circulation, 12*(1), e12000. https://doi.org/10.1002/pul2.12000

World Health Organization. (2020). *Global Cancer Observatory: Cancer today.* International Agency for Research on Cancer. Available from https://gco.iarc.fr/today/data/factsheets/cancers/15-Lung-fact-sheet.pdf

Witten, D., & Tibshirani, R. (2007). A comparison of fold-change and the t-statistic for microarray data analysis. *Analysis*, 1776, 58-85.

Yadav, S., & Shukla, S. (2016). Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. *2016 IEEE 6th International Conference on Advanced Computing (IACC),* 78-83. doi: 10.1109/IACC.2016.25

Zappa, C., & Mousa, S. A. (2016). Non-small cell lung cancer: Current treatment and future advances. *Translational Lung Cancer Research, 5*(3), 288–300. https://doi.org/10.21037/tlcr.2016.06.07

Zhang, L., Geisler, T., Ray, H., & Xie, Y. (2022). Improving logistic regression on the imbalanced data by a novel penalized log-likelihood function. *Journal of Applied Statistics, 49*(13), 3257-3277.