18th National and 3rd International Conference of Iranian Biophysical chemistry
هجدهمین همایش ملی و سومین همایش بین المللی بیوشیمی فیزیک ایران
25-26 Des, 2024, University of Hormozgan
۶-۵ دی ماه ۱۴۰۳، دانشگاه هرمزگان

Predicting osteoarthritis (OA) using deep convolutional neural network (DCNN) and transcriptome profile

Moslem Momen

Computer Engineering Department, Islamic Azad University - Bandar Abbas Branch, Hormozgan, Iran
moslemmomen97@gmail.com

Abstract

Osteoarthritis (OA) is a progressive joint disease characterized by cartilage degradation, bone remodeling, and inflammation, leading to pain and loss of mobility. Accurate prediction and early diagnosis of OA remain critical for effective intervention. Recent advancements in deep learning, particularly Deep Convolutional Neural Networks (DCNNs), have revolutionized the field of medical imaging by enabling precise pattern recognition in complex data such as MRI and X-ray scans. Additionally, transcriptome analysis provides valuable molecular insights into gene expression changes associated with OA progression. In this study, we attempted to predict OA by enriching DCNN-based models with transcriptome data to improve prediction accuracy and diagnosis. We hypothesized that by leveraging both DL techniques and gene expression molecular information, we could offer a comprehensive solution for identifying OA at its early stages and guiding personalized treatment strategies. Our results demonstrate that it is possible to accurately predict OA using gene expression data and deep neural networks, even with a limited sample size.

Key words: Osteoarthritis, Deep Convolutional Neural Networks, Transcriptome Analysis, Prediction Accuracy, Personalized Treatment Strategies

1. Introduction

Osteoarthritis (OA) is a prevalent chronic joint disease affecting over 300 million people worldwide, particularly in aging populations. Characterized by the gradual degradation of cartilage, bone remodeling, and inflammation, OA causes pain, stiffness, and mobility loss, greatly impacting patients' quality of life and creating a heavy healthcare burden [1]. Traditional OA diagnosis typically relies on symptom assessment and radiographic imaging; however, these methods often capture OA only in its later stages, which restricts the effectiveness of early interventions [2]. Early detection is vital to improve outcomes, yet current diagnostic tools lack the precision to identify subtle joint changes in OA's early stages.

Recent advancements in deep learning (DL), particularly in Deep Convolutional Neural Networks (DCNNs), have transformed medical imaging by enabling accurate identification of complex patterns in MRI and X-ray images [3]. Through their layered architecture, DCNNs can detect minute OA indicators that are difficult for human experts to discern, offering a more sensitive diagnostic tool [4]. Studies show DCNNs effectively recognize OA-related structural changes in joints, making them powerful tools for disease progression analysis [5, 6].

In addition to imaging advancements, transcriptomic analysis provides insights into the molecular changes underlying OA. By profiling gene expression, transcriptomics highlights disruptions linked to OA progression, including genes associated with inflammation and cartilage breakdown [7, 8]. When combined with imaging data, transcriptomics supports a comprehensive diagnostic approach, integrating anatomical and molecular information to enhance prediction accuracy.

This study proposes an innovative model combining DCNN-based imaging analysis with transcriptomic data to improve early OA detection and prediction accuracy. By integrating molecular data into DL models, we aim to bridge the gap between imaging and biological insights, supporting earlier detection and potentially informing personalized treatment strategies. We hypothesize that this combined approach will create a robust model that can better predict OA onset and progression, ultimately contributing to individualized patient care.

2. Methods

1.2 Metadata Curation and Processing

We utilized four gene expression datasets GSE117999, GSE51588, GSE57218, and GSE114007 downloaded from the Gene Expression Omnibus (GEO) database for our analysis. The GSE117999 dataset includes 24 samples, comprising 12 cartilage tissue samples from normal joints and 12 from osteoarthritis (OA) joints (Figure 1). GSE51588 contains 50 samples, with 10 subchondral bone tissue samples from normal joints and 40 from OA joints. The GSE57218 dataset consists of 73 samples, including 7 cartilage tissue samples from normal joints and 66 from OA joints. Finally, GSE114007 comprises 38 samples, with 18 cartilage tissue samples from normal joints and 20 from OA joints. These datasets provided a robust foundation for our subsequent differential gene expression analysis, allowing us to explore the molecular differences associated with OA.

Figure 1. A 2D medical illustration of an osteoarthritic knee joint showing characteristic damage associated with osteoarthritis. The image highlights cartilage degradation, cracks in the cartilage layer, and rough joint surfaces.

2.2 Differential Expression Analysis

To identify differentially expressed genes, we performed an RNA-seq data analysis using the edgeR package. The analysis began with the normalization of raw count data, followed by fitting a negative binomial generalized linear model to each gene to account for biological variability. We applied the likelihood ratio test (LRT) to determine the significance of differential expression between experimental conditions. The resulting statistics included log fold change (logFC), log counts per million (logCPM), likelihood ratio (LR), p-values, and false discovery rate (FDR) adjusted p-values [9].

For visualization, we created an MA plot using the ggplot2 package, where the log2 fold change was plotted against the average expression level (logCPM). Genes with significant differential

expression (adjusted p-value < 0.05) were highlighted: upregulated genes (logFC > 1) in green, downregulated genes (logFC < -1) in red, and non-significant genes (logFC between -1 and 1) in grey. To further focus on the most significant genes, we extracted those with both an adjusted p-value below 0.05 and an absolute logFC greater than 1. These genes were identified as the most relevant for downstream analyses, and their details, including gene names, logCPM, logFC, p-value, and FDR, were exported into a text file for further investigation and reporting [10].

3.2 Data modeling and DL models

We employed CNN neural network models to estimate predictive accuracy based on gene expression data from multiple studies (GSE51588, GSE117999, GSE57218, GSE114007) sourced from the GEO database. After preprocessing and aligning common genes across datasets, we split the data into training (GSE51588, GSE117999, GSE57218) and test (GSE114007) sets [11]. For analysis, we designed two neural network architectures: a CNN with 1D convolutional and max-pooling layers and an RNN with simple RNN and dense layers. Both models were trained for 20 epochs, using categorical crossentropy as the loss function. We evaluated model performance using AUC scores on the test set to measure their ability to distinguish between control and case conditions. This analysis demonstrated the predictive capabilities of CNNs and RNNs in recognizing gene expression patterns related to osteoarthritis.

3. Results and discussion

Figure 2 MA plot provides a visual representation of the differential expression of genes between osteoarthritis (OA) samples and control samples. Each point on the plot represents a gene, positioned based on two key metrics logCPM and log2FC.

A total of 1260 genes were identified as significantly associated with OA, either upregulated or downregulated. These genes provide insights into molecular pathways and biological processes potentially involved in OA development or progression, making them candidates for further study and validation. Shown in red, these genes are expressed at significantly higher levels in OA samples compared to controls, indicating potential roles in OA pathology. Shown in green, these genes have significantly lower expression in OA samples, suggesting they may be suppressed or less active in OA conditions. Represented in grey, these genes show minimal or no significant differential expression between OA and control samples, indicating a limited association with OA in this context.

Figure 2. MA Plot of Differential Gene Expression in OA vs. Control Samples, highlighting 1260 significantly upregulated (red) and downregulated (green) genes.

Our findings revealed that the simplest model outperformed the more complex architectures, achieving the highest AUC of 0.825. This model, which consisted of only two convolutional layers with 32 and 16 filters, demonstrated a better balance between model complexity and predictive performance. The fully complicated model, despite its depth and additional layers, yielded a slightly lower AUC of 0.814, while the medium complicated model performed the worst with an AUC of 0.612 (Figure 3). One of the most striking observations was the consistent difficulty across all models in accurately detecting true positive cases, as indicated by the low sensitivity scores. Despite achieving high specificity, which indicates the models capability to correctly identify non-cases, the low sensitivity suggests that the models struggled with true case identification. This trend may indicate potential overfitting in the more complex models, where the high number of layers and filters may have led to an excessive focus on the training data's nuances, reducing generalizability.

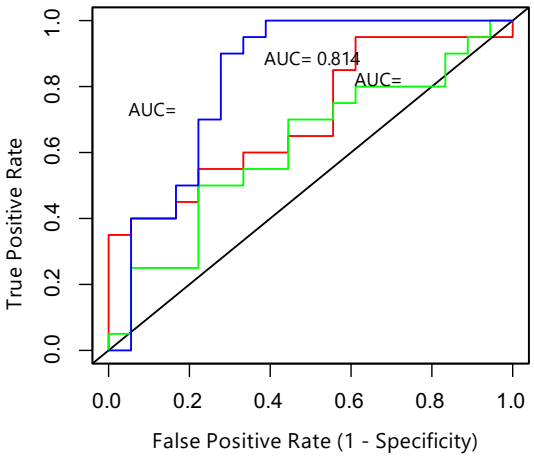

Figure 3. ROC Curve showing model performance for OA prediction: simplest model (AUC = 0.825, blue) outperformed complex (AUC = 0.814, red) and medium (AUC = 0.612, green) models.

These findings underscore the importance of model selection in the context of biomedical data, where simpler models may sometimes offer better performance, particularly in datasets with limited sample sizes or high-dimensional feature spaces. Our results align with previous studies suggesting that increasing model complexity does not always translate to improved performance and may, in some cases, lead to diminished returns due to overfitting. However, this study has some limitations. First, the dataset size, while focused on a specific set of genes related to osteoarthritis, may not capture the full spectrum of genetic variability associated with the disease. Additionally, the binary classification framework may have oversimplified the underlying biological complexity of osteoarthritis, which could benefit from a more nuanced multi-class or regression approach in future studies.

Future research could explore the use of transfer learning, where pre-trained models on larger datasets could be fine-tuned on osteoarthritis-specific data, potentially improving sensitivity. Additionally, integrating multi-omics data and clinical variables could enhance model accuracy and provide a more comprehensive understanding of the disease's genetic underpinnings. In conclusion, our study highlights the importance of balancing model complexity with performance metrics, particularly in the context of deep learning applications in genomics. While deep learning holds significant promise for advancing our understanding of complex diseases like osteoarthritis, careful consideration must be given to model architecture, especially when dealing with high-dimensional but small datasets.

References

[1] Hunter, D. J., & Bierma-Zeinstra, S. (2019). Osteoarthritis. The Lancet, 393(10182), 1745-1759.

[2] Bedson, J., & Croft, P. R. (2008). The discordance between clinical and radiographic knee osteoarthritis: a systematic search and summary of the literature. BMC Musculoskeletal Disorders, 9, 116.

[3] Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A., Ciompi, F., Ghafoorian, M., van der Laak, J. A., van Ginneken, B., & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. Medical Image Analysis, 42, 60-88.

[4] Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24-29.

[5] Tiulpin, A., Thevenot, J., Rahtu, E., Lehenkari, P., & Saarakkala, S. (2018). Automatic knee osteoarthritis diagnosis from plain radiographs: A deep learning-based approach. Scientific Reports, 8(1), 1-10.

[6] Ravì, D., Wong, C., Deligianni, F., Berthelot, M., Andreu-Perez, J., Lo, B., & Yang, G. Z. (2017). Deep learning for health informatics. IEEE Journal of Biomedical and Health Informatics, 21(1), 4-21.

[7] Lorenzo, P., Bay-Jensen, A. C., & He, Y. (2021). Molecular markers of osteoarthritis: current status and future perspectives. Clinical and Experimental Rheumatology, 39, 33-40.

[8] Yuan, X., Meng, H. Y., Wang, Y. C., Peng, J., Guo, Q. Y., Wang, A. Y., & Lu, S. B. (2020). Transcriptomic insights into osteoarthritis: From pathophysiology to drug development. Frontiers in Pharmacology, 11, 713.

[9] Robinson MD, McCarthy DJ, Smyth GK. "edgeR: a Bioconductor package for differential expression analysis of digital gene expression data." Bioinformatics. 2010 Jan;26(1):139-140. doi:10.1093/bioinformatics/btp616.

[10] Chen Y, Lun AT, Smyth GK. "Differential expression analysis of complex RNA-seq experiments using edgeR." Statistical Analysis of Next Generation Sequencing Data. 2014; 51-74. doi:10.1007/978-1-4939-3578-9_3.

[11] Sun, Yunchao, Hui Yang, Jiaquan Guo, Jian Du, Shoujiang Han, and Xinming Yang. "Identification of HTRA1, DPT and MXRA5 as potential biomarkers associated with osteoarthritis progression and immune infiltration." BMC Musculoskeletal Disorders 25, no. 1 (2024): 647.

[12] Zhang, Z., Zhao, Y., Liao, X., Shi, W., Li, K., Zou, Q. and Peng, S., 2019. Deep learning in omics: a survey and guideline. Briefings in functional genomics, 18(1), pp.41-57.