Can we envision a world where humans live without suffering from disease and fully enjoy their lives? How can we utilize AI to liberate humanity from illness? Having witnessed a family member suffer from cancer and diabetes, I realized how illness can physically and mentally limit an individual's freedom. This experience made these questions become the mission of my life. As such, I journeyed a consistent path at the intersection of machine learning and healthcare, from my graduate studies through VUNO Inc and Medical AI Co Ltd.

My research journey began during my **Master's program in Bioengineering at Seoul National University**, where I joined the Medical Biomechanics & Design Lab under Professor Jung Chan Lee in August 2015. My primary project was about developing a robot-assisted cardiopulmonary resuscitation (CPR) system. The research sought to optimize CPR by leveraging real-time bio-signal data. I contributed to identifying the optimal compression point for CPR based on end-tidal carbon dioxide measurements and echocardiography. Additionally, I developed a segmentation algorithm for the left ventricle in echocardiography using deep learning and proposed a three-dimensional left ventricular volume estimation method based on two-dimensional video data.

At the moment, we lacked sufficient labeled data to train machine learning models. But generating labels required assistance from cardiologists and significant manpower, which was impossible. To alleviate the problem, I learned echocardiography interpretation by myself from a cardiologist whom my advisor had introduced to me and labeled over 100 cases of echocardiography videos frame by frame. While time-consuming, this experience taught me the importance of labels and the difficulties of collecting them in machine learning projects.

Beyond the CPR project, I participated in diverse projects. I developed a system to estimate stroke volume from pulse pressure variation during swine hemorrhage experiments and supported the development of sensors for low-cost extracorporeal membrane oxygenation. These experiences helped me to have a thorough understanding of data collection systems and signal processing.

In August 2017, I joined **VUNO Inc**, an AI healthcare startup in Seoul, as a Research Scientist. These days, VUNO is considered a leading AI healthcare company in Korea, but when I joined the company, it was still in its early stages, having just completed Series B funding. Here, I took charge of research projects on developing predictive models using electrocardiograms (ECGs) and electronic health records (EHR).

One of my initial works was developing a model for early sepsis detection using EHR. I identified an optimal combination that improved neural network performance by 5% [1] and proposed a missing value imputation method using graphical relationships between clinical variables [2]. Another part of my work involved developing cardiovascular disease diagnosis models using ECGs. I developed a model to detect heart failure with reduced ejection fraction and myocardial infarction, and evaluated its effectiveness in a clinical setting [3-5]. For the myocardial infarction model, I designed an experimental setup comparing the performance of machine learning models to experienced cardiologists and proved the efficacy of an AI-based screening model [4].

From the summer of 2020, I concentrated on improving the robustness of machine learning models for bio-signals. Specifically, I explored self-supervised learning methods for ECGs using clinically important features [6] and proposed a temporal correlation regularization method for transformers to enhance their performance on time-series data [7].

At the same time, I studied data augmentation strategies for ECG classification models [8]. During predictive model development for electrocardiography (ECG), I often encountered performance degradation when testing on data with demographics different from the training dataset. I studied the mechanism that generates and records the heart's electrical signals to address this issue. Here, I realized that variations in heart size, position, and chest size—often influenced by demographic—affect ECG readings. Based on this, I hypothesized that simulating these variations through data augmentation could mitigate the distribution shift problem. Recognizing the interconnected nature of ECG leads and the joint

correlations among them, I developed a graphical model to represent these relationships. I then incorporated this model into data augmentation by perturbing the graphical structure of ECG leads to simulate demographic differences. The method I proposed resulted in performance improvements of about 3% across various datasets compared to existing methods.

After three years and seven months at VUNO Inc, I transitioned to **Medical AI Co Ltd** in April 2021 as a Senior Research Scientist. Here, my work focused on developing real-world healthcare solutions and addressing technical challenges such as distribution shifts and noisy labels.

One of my first projects was developing an Automated ML Experimentation system. I led a project to develop an automated ML system to streamline the process of building machine learning models. Given the growing demand for ML experiments from medical staffs and efficient model development, we set out to develop a system for machine learning experiments.

There were two factors to consider. The first factor was designing a system capable of performing a wide range of experiments by simply modifying configuration files that describe the experiment setting. Building a such system required careful consideration to ensure it works on a variety of training types, including supervised learning and self-supervised learning, as well as tasks like classification, regression, and segmentation. To minimize the mistake, I meticulously verified the implementation of all possible types of learning and integrated them by grouping commonalities and differences. The second factor was hyperparameter optimization. While many state-of-the-art methods are available as open-source these days, we had to identify the most suitable combination of network architecture, optimizer, and regularization techniques for specific tasks. For hyperparameter optimization, I chose to use the Ray framework for distributed parallel systems. Later, to enhance the efficiency of hyperparameter search, I conducted research on network architecture and data augmentations for ECG, reducing the hyperparameter search space of the system based on the insight from the research [9, 10]. This system was successfully deployed internally, and now even medical staff who lack expertise in ML can perform various ML experiments.

Between 2022 and 2023, I worked on training models for various Cardiovascular Disease Screening Products, including those for left ventricular systolic dysfunction, myocardial infarction, and aortic stenosis. Notably, AITIA-LVSD received recognition as an innovative medical device from the Korean FDA and won the prestigious UNIST-UCLA Digital Healthcare Challenge in 2023.

In parallel, I took the initiative on a project on Personalized Blood Pressure Estimation using ECG and photoplethysmography (PPG) signals, collaborating with Seoul National University Hospital (SNUH). We used real-world ICU data and found out the data included extreme noise and inaccurate labels. To address this, I developed signal and label filtering algorithms, the latter designed through close discussions with medical staff from SNUH. Additionally, I integrated test-time training to account for individual variability, achieving a 30% improvement in regression performance [11].

These projects posed numerous technical challenges, particularly related to distribution shifts in test data. To address this, I explored adversarial data augmentation (ADA), which simulates data distributions that the current model finds challenging. However, I found that existing ADA couldn't simulate temporal changes in ECG. Given the medical research indicating different temporal characteristics across demographic backgrounds, I believed addressing this was crucial to fully tackling the distribution shift problem. To solve this, I proposed differentiable time warping, a method that incorporates a time warping algorithm into ADA but leverages the frequency domain to overcome the non-differentiability of traditional time warping. This method tackled an out-of-distribution issue that the existing ADA failed to address, resulting in a 40% improvement in the F1 score on a particular dataset [12].

On the other hand**,** I also conducted research on inductive bias in medical data. Inductive bias refers to assumptions a learning algorithm makes, and incorporating appropriate inductive biases enables better generalization to unseen data. Unlike in the computer vision domain, one intriguing observation for ECG data was the effectiveness of convolution-based models, which often outperform pre-trained transformers

regardless of the size of the data. Thus, I tried using features of the convolution-baed model as guides for transformer blocks instead of self-supervised tasks, like contrastive learning and masked autoencoders. Specifically, I trained a transformer through a block-by-block knowledge transfer method, utilizing each block of a convolutional neural network to guide each transformer block. I observed that this makes the self-attention of the transformer acquire properties like translational invariance and locality. Additionally, transformers trained with these approaches outperformed convolution networks [13].

I wondered if I could train a transformer to exhibit convolutional characteristics by imposing specific constraints on self-attention rather than relying on a pre-trained convolutional network. By comparing the formulations of self-attention and convolution, I learned that under certain constraints, the attention matrix can correspond to a depth-wise convolution kernel. And the aggregation across heads with the following fully connected layer in the self-attention block corresponds to point-wise convolution. This perspective allows for the integration of convolution and self-attention into a single framework. Furthermore, by tuning coefficients that adjust the balance between self-attention and convolution, the model could effectively acquire properties of both self-attention and convolutional (translational invariance and locality). Experiments across various datasets demonstrated that the proposed method consistently outperformed both naive convolution-based and transformer networks by a considerable margin.

Developing products and conducting research on predictive models are enjoyable and fulfilling. However, a conversation with a doctor last year made me realize that many diseases we can now predict still lack definitive treatment. This left me frustrated, but, at the same time, it fueled my interest in exploring the fundamentals of disease and developing treatments that go beyond prediction. I have particularly developed an interest in bioinformatics.

As such, I am about to take a step to bioinformatics.

AI has already had a significant impact on bioinformatics. However, despite these advancements, technical challenges still persist in this field. For instance, bioinformatics data are collected under varied experimental conditions and methods, which can affect the quality and consistency of data. Thus, training models robust to variations arising in experiments is crucial. Setting appropriate structural inductive biases is also important, requiring insights that reflect the characteristics of different data types, especially when dealing with complex and high-dimensional data. Label noise is also prevalent, potentially causing models to learn incorrect patterns. Reducing their impact during training is essential. In fact, many of the challenges mirror those I faced when applying AI to the medical field, and I would like to use my experience to address these problems.

By addressing these technical challenges of applying AI to bioinformatics, I aim to further advance its application in this field. Based on this, I aspire to tackle a variety of bioinformatics challenges. I am particularly interested in enhancing therapeutic targeting at the gene to protein levels through the lens of AI. In relation to this, I hope to explore genomics and transcriptomics, which are essential for identifying key genetic drivers involved in disease. Additionally, I am interested in investigating structural biology, including protein-protein interactions and protein design, as these are important to both understanding diseases and developing effective treatments.

References

[1] **Byeong Tak Lee**, Kyung-Jae Cho, O-Yeon Kwon, Yeha Lee. Improving the Performance of a Neural Network for Early Prediction of Sepsis. 2019. Computing in Cardiology.

[2] **Byeong Tak Lee***, O-Yeon Kwon*, Hyungho Park, Kyung-Jae Cho, Joon-myoung Kwon, Yeha Lee. Graph Convolutional Networks-Based Noisy Data Imputation in Electronic Health Record. 2020. Critical Care Medicine.

[3] Jinwoo Cho*, **Byeong Tak Lee***, Joon-myoung Kwon, Yeha Lee, Hyunho Park, Byung-Hee Oh, Ki-Hyun Jeon, Jinsik Park, Kyung-Hee Kim. Artificial intelligence algorithm for screening heart failure with reduced ejection fraction using electrocardiography. 2021. ASAIO Journal.

[4] Changho Han, Youngjae Song, Hong-Seok Lim, Yunwon Tae, Jong-Hwan Jang, **Byeong Tak Lee**, Yeha Lee, Woong Bae, Dukyong Yoon. Automated Detection of Acute Myocardial Infarction Using Asynchronous Electrocardiogram Signals—Preview of Implementing Artificial Intelligence With Multichannel Electrocardiographs Obtained from Smartwatches: Retrospective Study. 2021. Journal of Medical Internet Research.

[5] **Byeong Tak Lee***, Joom-myoung Kwon*, Jinwoo Cho, Woong Bae, Hyunho Park, Won-Woo Seo, Iksung Cho, Yeha Lee, Jinsik Park, Byung-Hee Oh, Ki-Hyun Jeon. Usefulness of Deep Learning Algorithm for Detecting Acute Myocardial Infarction Using Electrocardiogram Alone in Patients With Chest Pain at Emergency Department: DAMI-ECG Study. 2023. Journal of Cardiovascular Intervention.

[6] **Byeong Tak Lee***, Seo Taek Kong*, Youngaje Song, Yeha Lee. Self-Supervised Learning with Electrocardiogram Delineation for Arrhythmia Detection. 2021. Annual International Conference of the IEEE Engineering in Medicine & Biology Society.

[7] Kyung Geun Kim*, **Byeong Tak Lee***. Self Attention with Temporal Prior: Can We Learn More from Arrow of Time?. 2024. Frontiers in Artificial Intelligence.

[8] Kyung Geun Kim*, **Byeong Tak Lee***. Graph Structure Based Data Augmentation Method. 2024. Biomedical Engineering Letters.

[9] **Byeong Tak Lee***, Yong-Yeon Jo*, Seon-Yu Lim, Youngjae Song, Joon-myoung Kwon. Efficient Data Augmentation Policy for Electrocardiograms. 2022. Proceedings of the 31st ACM International Conference on Information & Knowledge Management.

[10] **Byeong Tak Lee**, Yong-Yeon Jo, Joon-myoung Kwon. Revisiting Neural Network Scale for ECG Classification. (under review)

[11] Yong-Yeon Jo*, **Byeong Tak Lee***, Beom Joon Kim, Jeong-Ho Hong, Hak Seung Lee, Joon-myoung Kwon. New Test-Time Paradigm for Real-World Biosignal: Concept and Its Approach. (under review)

[12] **Byeong Tak Lee**, Joon-myoung Kwon, Yong-Yeon Jo. TADA: Temporal Adversarial Data Augmentation for Time Series Data. (under review)

[13] **Byeong Tak Lee**, Yong-Yeon Jo, Joon-myoung Kwon. On the Inductive Bias Transfer with Knowledge Distillation for Real World Data. 2022. Workshop on Applied Machine Learning Methods for Time Series forecasting.