의학석사 학위논문

손목 x-선을 활용한 미숙아 대사성 골 질환 진단 딥러닝 모델 구축

Developing a deep learning model to diagnose metabolic bone disease of prematurity using wrist x-ray results of preterm infants

2022년 8월

서울대학교 대학원

의학과 소아과학 전공

Seul Gi Park

손목 x-선을 활용한 미숙아 대사성 골 질환 진단 딥러닝 모델 구축

Developing a deep learning model to diagnose metabolic bone disease of prematurity using wrist x-ray results of preterm infants

지도 교수 김 이 경

이 논문을 의학석사 학위논문으로 제출함

2022년 4월

서울대학교 대학원

의학과 소아과학 전공

Seul Gi Park

Seul Gi Park의 의학석사 학위논문을 인준함

2022년 7월

위 원 장 김 한 석 (인)

부위원장 김 이 경 (인)

위 원 천 정 은 (인)

ABSTRACT

Developing a deep learning model to diagnose metabolic bone disease of prematurity using wrist x-ray results of preterm infants

Seul Gi Park

Medicine, Pediatrics

The Graduate School

Seoul National University

Background: Metabolic bone disease (MBD) of prematurity is an important complication of prematurity and accurate diagnosis and timely intervention should be made for preterm infants.

Objective: To develop a diagnostic tool for MBD of prematurity via deep learning by using wrist x-rays of preterm infants.

Methods: Study enrolled preterm infants whose birth weight was less than 1500g born at Seoul National University Children's Hospital and admitted to Neonatal Intensive Care Unit from 2010 to 2020. Demographic and clinical information as well as wrist x-rays taken between 4-8 weeks of postnatal age were collected retrospectively. Two types of regions of interests ('ROI 0' and 'ROI 1') were annotated for deep learning model training. Demographic and clinical data was analyzed to determine the factors associated with MBD of prematurity, thus evaluating the representativeness of our study population. Wrist x-ray images were used to train and develop a diagnostic model via various deep learning algorithms, including AlexNet, DenseNet-121, ResNet-50, ResNext-50, VGG-19, CheXNet, and EfficientNet-b3.

Results: Fourteen percent (116/814) of enrolled patients were diagnosed with MBD of prematurity between 4–8 weeks of postnatal age. Analysis of clinical information revealed that birth weight less than 1000g (82.8% vs. 37.5%, p<0.001), gestational age less than 28 weeks (75.0% vs. 29.5%, p<0.001), parenteral nutrition longer than or equal to 28 days (49.1% vs, 12.0%, p<0.001) were statistically significant risk factors of MBD of prematurity. These risk factors concurred with renowned risk factors of MBD, suggesting that our population could represent general preterm population and our ground truth is reliable. Deep learning models developed by EfficientNet-b3 and VGG-19 using 'ROI 0' appeared to show the best quality of performance demonstrated by highest F1-score (0.844 for both models) and AUROC (0.962 for EfficientNet-b3 and 0.968 for VGG-19). 'ROI 0' EfficientNet-b3 model and VGG-19 model both showed sensitivity of 0.907, specificity of 0.924, positive predictive value of 0.790, negative predictive value of 0.969, and accuracy of 0.915.

Conclusion: Novel deep learning models to diagnose MBD of prematurity have been developed as a result. Our models showed sensitivity of 0.907, specificity of 0.924, and accuracy of 0.915. If applied to clinical settings, it would assist clinicians, especially for those who are novice, to detect MBD more accurately and conveniently, thereby enabling timely management to treat and prevent disease progression for preterm infants.

.....

Keywords: metabolic bone disease of prematurity, artificial intelligence, deep learning, prematurity, wrist x-rays
Student Number: 2020-29040

CONTENTS

Abstract i
Contents iii
List of tables and figures iv
List of Abbreviations v

Introduction 1
Material and methods 3
Results 8
Discussion 12
Conclusion 15

References 27
Abstract in Korean 29

LIST OF TABLES AND FIGURES

Figure 1. Wrist X-ray Images of Normal and MBD.....16
Figure 2. Image Annotation – Region of Interest 17
Figure 3. Image Augmentation 18
Figure 4. Flow Diagram – Cohort 19
Table 1. Demographic and Clinical Characteristics.....20
Table 2. Multivariable Logistic Regression of Clinical Data.....21
Table 3. Risk factors of MBD 22
Figure 5. Deep Learning Flowchart 23
Table 4. Performance Metrics of Designed Models.....24
Figure 6. ROC curves of Designed Models.....25
Figure 7. Gradcam Images.....26

LIST OF ABBREVIATIONS

MBD	Metabolic bone disease
AI	Artificial intelligence
SNUCH	Seoul National University Children's Hospital
NICU	Neonatal Intensive Care Unit
ALP	Alkaline phosphatase
VLBW	Very low birth weight
ELBW	Extremely low birth weight
GA	Gestational age
TPN	Total parenteral nutrition
EN	Enteral nutrition
EMR	Electronic Medical Records
CNN	Convolutional neural networks
DICOM	Digital Imaging and Communications in Medicine
PNG	Portable Network Graphics
ROI	Region of interest
CPU	Central Processing Unit
GPU	Graphics Processing Unit
ROC	Receiver operating characteristics
AUC	Area under curve
AUROC	Area under ROC
PPV	Positive predictive value
NPV	Negative predictive value
TP	True positive
TN	True negative
FP	False positive
FN	False negative
IV	Intravenous

INTRODUCTION

Metabolic bone disease (MBD) of prematurity is also known as osteopenia or rickets of prematurity.[1] It is prevalent in preterm infants especially whose gestational age is younger than 28 weeks. Among preterm infants whose birth weight is below 1500g (very low birth weight, VLBW) and below 1000g (extremely low birth weight, ELBW), about 16–40% are diagnosed with MBD of prematurity with the peak incidence at the postnatal age of 4–8 weeks.[1–3] It is known to be attributable to calcium and phosphorus deficiency due to decreased intake or absorption in most cases.[1,4,5,6]

Exact incidence of MBD of prematurity among preterm infants is still unknown partly due to the lack of consensus regarding the diagnostic definition of this disease.[7] Therefore, several screening protocols exist instead, to select and evaluate infants who are at high risk of MBD of prematurity by using their biochemical markers and wrist x-rays.[7] Clinicians suspect the presence of MBD if the serum levels of phosphorus and calcium are persistently low, while serum alkaline phosphatase (ALP) level is increasing.[4,8] For more definitive diagnosis, wrist x-ray is often obtained which might demonstrate the cupping or fraying at radius metaphysis, the classical radiological findings of MBD of prematurity (Figure 1).[3,7,9,10] However, such radiological findings are known to appear only after bone mineralization has been reduced by 20–40% and identifying early stage of MBD on x-ray images is challenging.[7]

In recent years, numerous articles about artificial intelligence (AI) revealed that hundreds of applications of deep learning to medical images have made that offered opportunities to improve the speed, accuracy, and quality of image interpretation and radiological diagnosis.[11,12] Since MBD of prematurity is one of many major complications of prematurity that should be prevented for better long term outcomes, this study aimed to develop a deep learning program for MBD of prematurity by using wrist x-rays obtained between 4–

8 weeks of postnatal ages among preterm infants born less than 1500g. Programmed diagnostic algorithm is expected to enhance the accuracy and efficacy of MBD diagnosis for preterm infants in clinical settings, thereby allowing earlier interventions and treatments to prevent disease progression.

MATERIAL AND METHODS

Study Population and Setting

This study is designed as a retrospective study including preterm infants weighed less than 1500g at birth and were born at Seoul National University Children's Hospital (SNUCH) from 2010 to 2020 and admitted to neonatal intensive care unit (NICU) of SNUCH.

Patients were excluded if discharged, transferred to other hospitals, or died before 28 days of life, if malformations, chromosomal abnormalities, or metabolopathies were associated, or if there were no wrist x-rays taken between 4-8 weeks of life (25-59 postnatal days). While reviewing the collection of wrist x-ray images of enrolled patients, few images difficult to be read correctly were excluded from the study, especially those that had intravenous catheter line positioned at metaphysis area and those that showed overlapped radius and ulna due to the improper position of patients.

Data Collection and Interpretation

Department of neonatology, department of radiology, and transdisciplinary department of medicine and advanced technology cooperated throughout the study in data collection, interpretation, and development of a deep learning model.

First, neonatologists from SNUCH collected demographic and basic clinical information, including gestational age (GA), birth weight, birth weight percentage, gender, associated anomalies, duration of total parenteral nutrition (TPN) in days, and days taken to reach enteral nutrition (EN) full feeding (≥100mL/kg/day), by reviewing their electronic medical records (EMR). Wrist x-rays taken between 4-8 weeks of life (25-59 days, giving a window period of 3 days) were collected. Data from SNUCH was utilized as an internal dataset.

Collected wrist x-rays were reviewed by professional pediatric radiologist (C.J.E. with 22 years of post-fellowship

experience) and neonatologist (K.E.K. with 19 years of post-fellowship experience) who were blinded by clinical or laboratory conditions of patients. Images were denoted as '0' if normal, and '1' if diagnosed with MBD of prematurity. Images with an indeterminate of discrepant categorization were jointly reviewed and discussed, then categorized by consensus.

Labelled images were used as ground truth,[13] and transdisciplinary department of medicine and advanced technology team utilized this set of data to develop a deep learning model.

Imaging Dataset Partitioning

Convolutional neural network (CNN) is a complex computational model using multiple algorithm layers to achieve high-level interpretations of data. This is currently applied widely for classifying medical images,[14] generally based on a supervised approach.[15] For deep learning model training and testing, wrist x-ray images labeled by professional radiologists were used as a defined ground truth.

Considering the incidence of MBD of prematurity, collected raw image dataset from SNUCH showed a considerable imbalance between the number of normal images and the number of MBD images. Number of normal images outweighed that of MBD images, thus the majority class went through undersampling.

After undersampling of normal images, the final imaging dataset was split into training set, validation (tuning) set, and test set at 7:1:2 ratio. Images were split via stratified random sampling in order to reduce the potential for an uneven or nonrandom distribution of normal and MBD images. Images belonged to the same patients were stratified into the same set to minimize potential bias.

Data Preprocessing

Wrist images were converted from Digital Imaging and Communications in Medicine (DICOM) to Portable Network Graphics

(PNG) images by using Python version 3.7.11.

Annotation was performed for each image by drawing a labeled bounding box using the software tool roLabelImg version 1.8.0. For each image, two region of interest (ROI) boxes were drawn. Directions of drawing 'ROI 0' and 'ROI 1' are described in Figure 2. As defined by Koo et al., metaphyseal alterations of radius are the key features determining the presence of MBD.[10] Thus, 'ROI 0' was drawn as a square surrounding the metaphyseal plate of radius. 'ROI 1' was drawn as a rectangle, which was extended downward from 'ROI 0' to cover more of trabecular bones of radius. Trabecular bones are known to respond to metabolic changes faster than cortical bones and changes are most prominent in the ends of the long tubular bones, particularly in the distal radius that have a relatively large proportion of trabecular bone.[16] Annotation data were saved to a single XML file with designated labels corresponded to the categorization of either normal or MBD of prematurity. ROIs were cropped then resized to a resolution of 300x300 matrix for EfficientNet-b3 and 224x224 matrix for other deep learning algorithms. All imaging data underwent normalization process using ImagNet.

To overcome the volume discrepancy between normal images and MBD images, MBD images of the training data sets were augmented by various techniques; such as, random rotations between ±15 degrees, random brightness contrast, GaussianBlur, and Gauss Noise methods, resulting in 3 times the image augmentation (Figure 3). All types of augmentation were performed using Albumentation version 1.1.0 (https://albumentations.ai/docs/api_reference/augmentations/transforms/). Rotation was made by 1 degree within the range between −15 and +15 degrees. Random brightness and contrast were both adjusted by −0.2 to +0.2 using default settings of Albumentation. For Gaussian Blur technique, a random sized kernel with center pixel of (3,7) was applied to obtain the new values for the center pixel of (3,3), (4,4), (5,5), (6,6), and (7,7). GuassNoise

method was done with variance range for noise having minimum and maximum values of 10 and 50, respectively.

Model Development and Testing

In this study, we experimented with 7 algorithms: AlexNet[17], DenseNet121[18], ResNet50[19], ResNext50[20], ChexNet[21], VGG19[22], EfficientNet-B3[23]. ImageNet-pretrained models were used as an initial parameter for algorithm training. All models were trained with the same hyper-parameters, such as Adam optimizer (learning rate: 1e-4), epochs (early stopping), batch size (8), and input size of 300x300 for EfficientNet-B3 and 224x224 for other algorithms.

All image processing and CNN development work were performed on central processing unit (CPU) and graphics processing unit (GPU) nodes composed of the Intel Core i5-11400F CPU and NVIDIA RTX 3060 (12GB). All coding was performed using the Pytorch[24] deep learning platform.

Efficiency of developed diagnostic models using aforementioned neural network architectures as well as 'ROI 0' and 'ROI 1' were determined individually, then the most efficient model would be selected for further programming.

Data Analysis

Patients were grouped into normal and MBD group. If patients had no evidence of MBD on wrist x-ray during 4-8 weeks of life, they were grouped as normal. On the other hand, when patients were diagnosed with MBD on wrist x-ray at least once during the study period, they were grouped as MBD.

Statistical comparison of demographic features and clinical information between the groups were made by SPSS IBM Statistics 26.0. Chi-square test was used for categorical variables and student's t-test was used for continuous variables. To reduce confounding effects of variables, multivariable logistic regression analysis was performed and adjustment for sex, gestational age, birth

weight, TPN period, full EN reaching period, length of hospital stays, 1-minute Apgar score, and 5-minute Apgar score were made. Statistical significance was defined as $p<0.05$.

For deep learning model performance evaluation, ROC with area under curve (AUC) was generated to define test accuracy. Evaluation metrics; such as, accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F1-score were calculated.[25]

Ethics Statement

This study protocol was reviewed and approved by the Institutional Review Board (IRB) of Seoul National University Hospital (IRB No. H-2104-056-1210). Requirement for informed consent was waived due to the retrospective design of the study.

RESULTS

Study Population

During 2010–2020, 1110 preterm infants weighed less than 1500g at birth were born at SNUCH and admitted to SNUCH NICU. Among these infants, 296 patients were excluded, in which 257 infants were discharged, transferred to other hospitals or died before 28 days of life, and 39 infants either did not take wrist x-rays between 4–8 weeks of postnatal age or did not have qualifying images. None had major congenital anomalies, metabolopathies, or chromosomal abnormalities. Total of 814 patients were enrolled and 14.3% (116/814) were diagnosed radiologically with MBD of prematurity at least once between 4–8 weeks of life (Figure 4).

Among this population, the number of wrist x-ray files taken between 4–8 weeks of postnatal age was 1134, containing 2239 wrist images. Of these, 114 files (271 images) were MBD images and remaining 990 files (1968 images) were categorized as normal.

Demographic and Clinical Data of Study Population

814 preterm infants were finally enrolled in the study population in which 416 (51.1%) were male and 398 (48.9%) were female. Average values for our study population were summarized in Table 1. Gestational age and birth weight were 28.97 weeks and 1033.48 grams, respectively. 18.47 days of TPN was given on average, while it took 16.01 days to reach full enteral feeding. Patients were admitted in NICU for 72.55 days on average and their mean Apgar score was 4.33 at 1 minute and 6.80 at 5 minutes.

Enrolled patients were divided into two groups; MBD group or normal group. Patients were grouped as MBD if radiological diagnosis of MBD has been made at least once during the study period. Of 814 patients, 116 (14.3%) were classified as MBD and 698 (85.7%) belonged to normal group.

Since infants born at lower gestational age tend to be

associated with lower birth weight and these infants might also be associated with longer TPN period, longer duration to reach full enteral feeding, longer hospital days, and lower Apgar scores. Considering such confounding effects among the variables, multivariate logistic regression analysis was performed, in which adjustment for sex, GA, birth weight, TPN period, full EN reaching period, length of hospital stays, and Apgar scores was made (Table 2). After adjustment, MBD group and normal group showed significant differences in gestational ages, birth weight, TPN period, and length of hospital stays. Lower GA (OR 0.751, 95% CI 0.658−0.857, p=0.000), lower birth weight (OR 0.997, 95% CI 0.996−0.999, p=0.000), longer lengths of TPN period (OR 1.022, 95% CI 1.001−1.044, p=0.040), and longer hospital stays (OR 1.007, 95% CI 1.000−1.015, p=0.046) were associated with higher risk of MBD or prematurity.

Risk Factors of MBD of prematurity

Small for gestational age, birth weight less than 1000g, gestational age less than 28 weeks, and TPN period longer than 28 days are widely recognized risk factors of MBD,[1,4,26,27] as also suggested from our data. The risk factors were analyzed by chi−square tests and birth weight less than 1000g (82.8% vs. 37.5%, p<0.001), GA less than 28 weeks (75.0% vs. 29.5%, p<0.001), TPN longer than or equal to 28 days (49.1% vs, 12.0%, p<0.001) showed statistical significance in association with MBD of prematurity (Table 3).

Deep Learning Flow Using Internal Data

Total number of normal images was 1968 while MBD images were only 271. For proper training, such data size discrepancy in between normal and MBD groups was overcome by downsampling the majority class, the normal group. After randomly performing downsampling of the normal group, final internal dataset contained

833 normal images from 298 patients and 271 MBD images from 116 patients. This final dataset was divided into training set, validation (tuning) set, and test set at 7:1:2 ratio by stratified random sampling. To overcome the insufficient number of MBD images available, imaging augmentation was performed which tripled the number of MBD images. Numerous types of deep learning algorithms, such as AlexNet[17], DenseNet121[18], ResNet50[19], ResNext50[20], ChexNet[21], VGG19[22], EfficientNet-B3[23], were applied to develop our model and the performance metrics of developed models were evaluated. The flowchart of deep learning process is summarized in Figure 5.

Performance Evaluation

Various models to diagnose MBD of prematurity based on wrist x-rays were created by using both 'ROI 0' and 'ROI 1'. Calculated performance metrics and ROC curves of each model are demonstrated in Table 4 and Figure 6.

Overall, the models developed by EfficientNet-b3 and VGG-19 using 'ROI 0' appeared to show the highest quality of performance, demonstrated by the highest F1-score. 'ROI 0' model by EfficientNet-b3 revealed the sensitivity of 0.907, specificity of 0.924, PPV of 0.790, NPV 0.969, F1-score of 0.844, accuracy of 0.915, and AUROC of 0.962. 'ROI 0' model created by VGG-19 showed the sensitivity of 0.907, specificity of 0.924, PPV of 0.790, NPV 0.969, F1-score of 0.844, accuracy of 0.915, and AUROC of 0.968, presenting similar performance efficacy.

The test dataset for 'ROI 0' EfficientNet-b3 model contained 171 normal images and 54 images. Our model made correct diagnosis for 92.4% of normal images (158/171, true negative, TN, specificity) and 90.7% of MBD images (49/54, true positive, TP, sensitivity). Reviewing the Gradcam images of TP and TN cases suggested that the model was focusing mainly on the area of metaphysis of radius as we intended. This model also made a few incorrect diagnoses, where 7.6% (13/171, false positive, FP) of

normal images were classified as MBD and 9.5% (5/54, false negative, FN) of MBD images were named normal. From the Gradcam images of FP and FN, it was noticed that the model focused less strongly on the metaphysis area of radius or focused on ulna or intravenous (IV) fluid line instead of radius. Figure 7 shows Gradcam images.

DISCUSSION

MBD of prematurity is an important complication of prematurity which could even result in pathological fractures if left untreated.[2,5] Thus in order to make MBD diagnosis more efficiently and accurately in clinical settings, even for those who lack clinical experience and for those who could not be supported by professional radiologists, we aimed to create a deep learning model to diagnose MBD of prematurity. As a result, novel diagnostic deep learning models for MBD of prematurity were developed showing sensitivity of 0.907 and specificity of 0.924.

Data was collected from January 2010 to December 2020 at SNUCH NICU. All included patients were very low birth weight infants (VLBW) whose birth weight was less than 1500g. Considering the fact that the peak incidence of MBD of prematurity occurs at 4-8 weeks of postnatal age, their wrist x-ray data from this period was collected for each patient. At SNUCH NICU, weekly blood sampling and monthly wrist x-rays are taken according to our center's MBD screening protocol. Thus, it was possible to collect feasible amount of wrist x-ray data.

Finally included patients were 814 patients and 14.3% was diagnosed with MBD on their wrist x-ray during 4-8 weeks of life. The observed incidence of MBD in our study population was close to the reported incidence of MBD of prematurity, about 16-40% among VLBW and ELBW infants.[2] Analysis of clinical findings and risk factors of our study population proved that lower GA, lower birth weight, and longer TPN period were associated with higher risk of MBD, which also concurred with the known risk factors of MBD of prematurity. These statistical findings strongly propose that our study population was appropriately representing general preterm population and that our ground truth is reliable.

After undersampling of majority class (normal images) then augmentation of minority class (MBD images), the final imaging

dataset was selected. Training set composed of 585 normal images and 573 MBD images and tuning set consisted of 77 normal images and 78 rickets images. Each images were annotated with both 'ROI 0' surrounding the metaphysis of radius and ulna and 'ROI 1' including metaphasis and trabecular bone of radius. Then the final dataset was applied to various deep learning algorithms to train and develop our deep learning diagnostic model.

The models developed by EfficientNet-b3 and VGG-19 using 'ROI 0' showed the best quality of performance. 'ROI 0' EfficientNet-b3 model showed sensitivity: 0.907, specificity: 0.924, PPV: 0.790, NPV: 0.969, F1-score: 0.844, accuracy: 0.915, AUROC: 0.962. When tested using 171 normal images and 54 images, our model made correct diagnosis for 92.4% of normal and 90.7% of MBD cases. Reviewing the Gradcam images, our model mostly detected metaphysis area of radius correctly. However, in some cases, the program focused more on ulna metaphysis instead of radius. This might have decreased the accuracy of our model since radiological definition of MBD of prematurity should not be made for isolated cupping of distal ulna.[28] Other Gradcam images focused on artifacts, including IV catheter line placed on hands. Although this might have influenced the performance of the program, we decided to keep those images with artifacts in imaging dataset in order to reflect real clinical settings. It is essential to keep IV catheter lines for preterm infants to process medical practices in NICU. Also, it is nearly impossible to control the spontaneous movements of infants while taking wrist x-rays in real settings. The final model was developed after including such images containing artifacts in the dataset, so it could be more suitable in real clinical settings.

As the field of AI offered opportunities to improve the medical imaging interpretation, Meda et al. conducted a study already to establish AI diagnostic tool for MBD.[29] This study used 104 MBD images and 264 normal medical images collected from patients younger than 7 years old. Until now, no researches have been made

to develop such model specifically for preterm infants. To the best of our knowledge, our study is the first study developing a deep learning model for MBD diagnosis among preterm infants. Unlike previous studies regarding the AI application to diagnose MBD of prematurity, our study used two different regions of interests in order to find the most ideal diagnostic algorithms. Also, the size of our imaging dataset was larger than other similarly designed studies and was obtained solely from the preterm population.

Upon the completion of our model development, it was expected to help clinicians to diagnose MBD of prematurity more easily. We believe that the model would aid novice clinicians and clinicians working without a support from radiologist to detect the presence of MBD, so earlier intervention could be made to prevent disease progression.

There also exist a few limitations. Deep learning is a data hunger program that always prefers bigger size of dataset. Thus, the performance of our model could be advanced if bigger dataset was available. Our dataset was collected retrospectively from a single center, so its generalizability to general preterm population is limited. Also, current model was developed to only detect whether the image is MBD or normal. It was not trained to differentiate the status of MBD, whether the disease is in its early stage or in healing state. Because the study was conducted in a single center, only internal dataset was available for training and testing. In order for our model to be applicable to general preterm infants, performing external validation test and user study to evaluate the efficacy should be considered.

Serum levels of ALP, calcium, and phosphorus are most commonly used biochemical markers to determine infants who are at higher risk of MBD.[2,3] Among these markers, high ALP level is the most well-known reliable biomarker suggesting the presence of MBD of prematurity and several studies already have reported the association between ALP levels and the radiological diagnosis of

MBD.[30] Further studies incorporating the biochemical markers to the developed model could be considered so that more practical and sophisticated diagnostic model to be created.

Moreover, based on the diagnostic algorithm developed from this study, a subsequent study developing a new screening deep learning model for MBD of prematurity by using infantogram could also be considered. Since infantograms are taken routinely during NICU admission for clinical practices, if such model is developed, it could also decrease the number of wrist x-rays being taken, reducing the radiation exposure hazard for preterm infants.

CONCLUSION

A novel diagnostic deep learning model for MBD of prematurity has been created having sensitivity of 0.907 and specificity of 0.924 at the end of the study. Further conduction of external validation and user study could enhance the quality of the developed model. This model would assist clinicians to detect MBD more accurately and conveniently, thereby enabling timely management to treat and prevent disease progression for preterm infants.

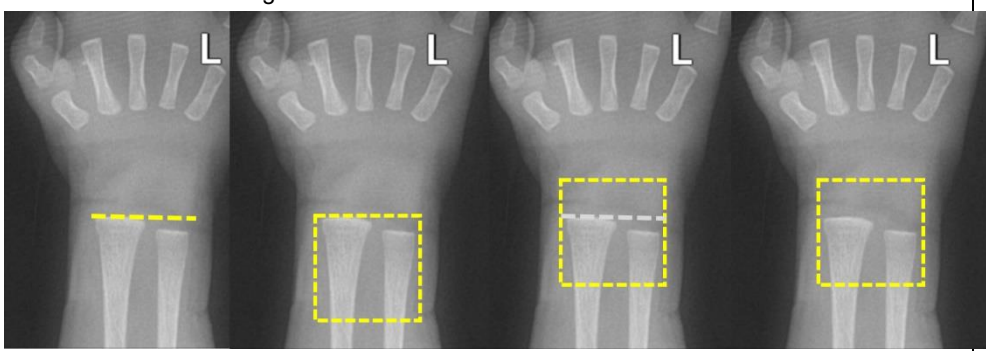
Figure 1. Wrist X-ray Images of Normal and MBD

Normal	MBD
	

Figure 2. Image Annotation - Region of Interest

ROI 0

(1) Draw a straight line on the metaphysis of radius, perpendicular to the y-axis of radius. Length of the line should cover the metaphysis of both radius and ulna.
(2) Create a square box having the line drawn above as its side.
(3) Adjust the box position to locate the metaphysis of radius at upper 1/3 (lower 2/3) of the box's height.



ROI 1

(1) Draw another 'ROI 0' square around the metaphysis of radius and ulna as mentioned above.
(2) Extent the length of the square to create a rectangle. Length of the rectangle should be 1.5 times longer than width.
(3) Position the box to locate its top side as same as ROI 0.

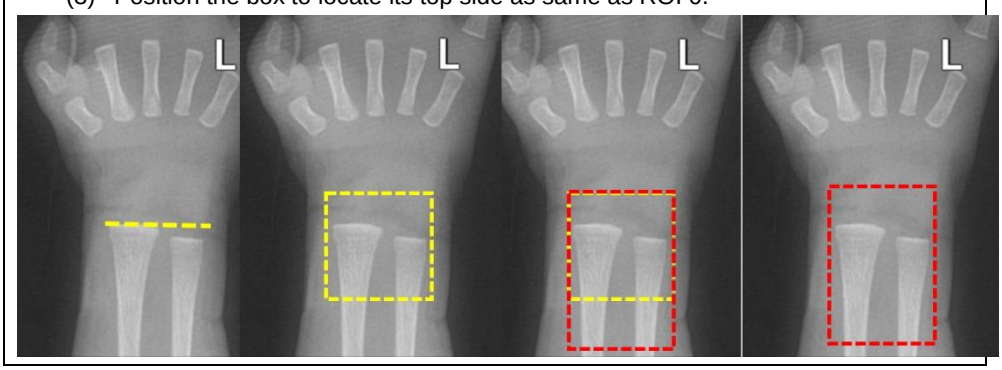


Figure 3. Image Augmentation

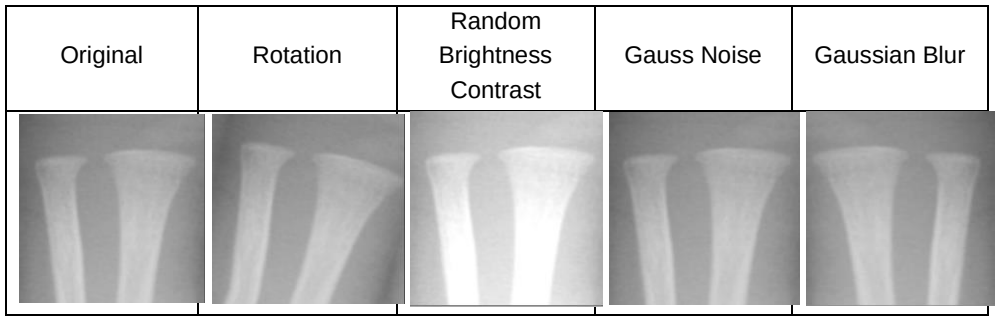


Figure 4. Flow Diagram - Cohort

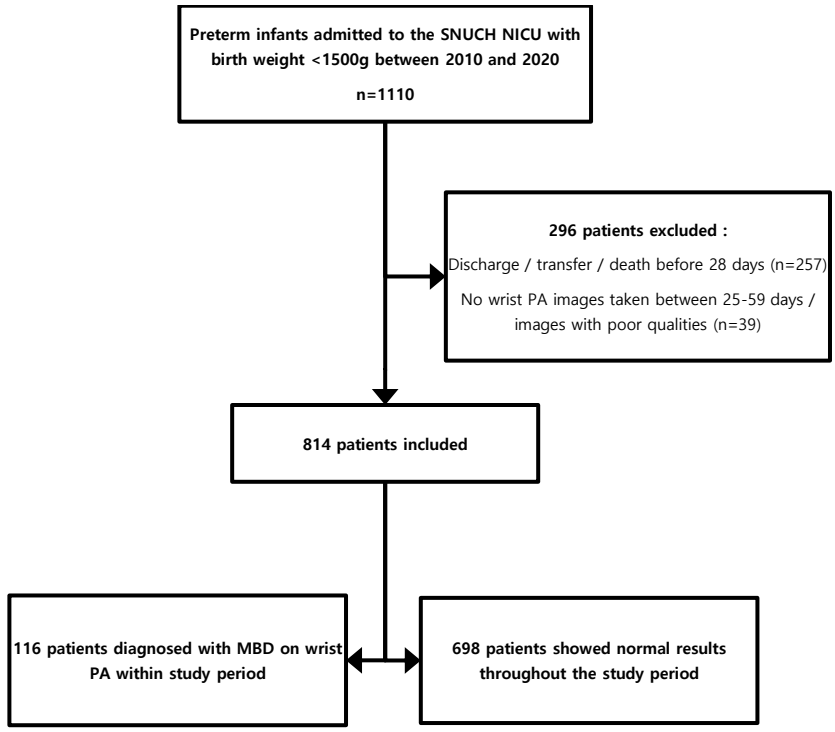


Table 1. Demographic and Clinical Characteristics

Table 1. Demographic and clinical characteristics	
	Total (n=814)
Sex (n, %)	
Male	**416 (51.1)**
Female	**398 (48.9)**
GA (weeks)	
Average	**28.97**
Birth weight (g)	
Average	**1033.48**
TPN period (days)	
Average	**18.47**
EN ≥100mL/kg/d (days)	
Average	**16.01**
Length of stay (days)	
Average	**72.55**
1 minute Apgar	
Average	**4.33**
5 minutes Apgar	
Average	**6.80**

Table 2. Multivariable Logistic Regression of Clinical Data

Table 2. Multivariable logistic regression analysis for MBD of prematurity

Variables	MBD of prematurity		
	OR	95% CI	P-value
Sex (Female)	0.643	0.392-1.053	0.079
GA (weeks in days)	0.751	0.658-0.857	<0.001
Birth weight (g)	0.997	0.996-0.999	<0.001
TPN period (days)	1.022	1.001-1.044	0.040
EN ≥100mL/kg/d (days)	0.997	0.972-1.022	0.796
Length of stay (days)	1.007	1.000-1.015	0.046
1 minute Apgar	1.204	0.988-1.468	0.065
5 minutes Apgar	0.999	0.891-1.218	0.993

OR, odds ratio ; CI, confidence interval
P-values are calculated by the analysis of multivariable logistic regression model for MBD of prematurity, after adjustment for sex, gestational age, birth weight, TPN period, full EN reaching period, length of stay, 1 minute and 5 minute Apgar scores.

Table 3. Risk Factors of MBD

Table 4. MBD risk factors			
	MBD (n=116)	Normal (n=698)	P-value
SGA_10p (n, %)			
<10 percentile	30 (25.9)	177 (25.4)	0.908
≥10 percentile	86 (74.1)	521 (74.6)	
SGA_3p (n, %)			
<3 percentile	20 (17.2)	101 (14.5)	0.437
≥3 percentile	96 (82.8)	597 (85.5)	
Birth weight (n, %)			
<1000g	96 (82.8)	262 (37.5)	<0.001
≥1000g	20 (17.2)	436 (62.5)	
GA (n, %)			
< 28 weeks	87 (75.0)	206 (29.5)	<0.001
≥ 28 weeks	29 (25.0)	492 (70.5)	
TPN period (n, %)			
≥ 28 days	57 (49.1)	84 (12.0)	<0.001
< 28 days	59 (50.9)	614 (88.0)	

Figure 5. Deep Learning Flowchart

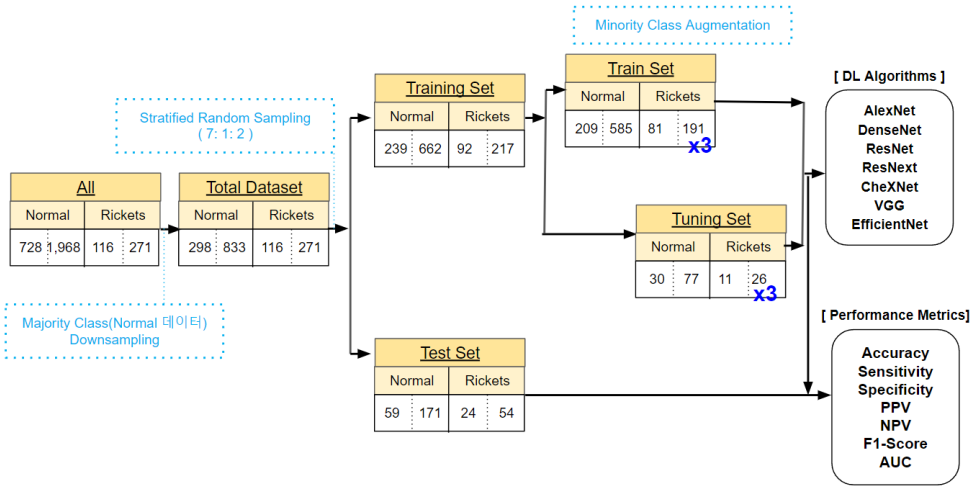


Table 4. Performance Metrics of Designed Models

ROI	DL Algorithm	Sensitivity	Specificity	PPV	NPV	F1-score	Balanced Accuracy	AUROC
0	AlexNet	0.870	0.912	0.758	0.957	0.810	0.891	0.946
	DenseNet-121	0.703	0.865	0.623	0.902	0.660	0.784	0.872
	ResNet-50	0.870	0.935	0.810	0.958	0.839	0.903	0.961
	ResNext-50	0.851	0.912	0.754	0.951	0.800	0.882	0.955
	ChexNet	0.925	0.894	0.735	0.974	0.819	0.910	0.966
	EfficientNet-b3	0.907	0.924	0.790	0.969	**0.844**	0.915	**0.962**
	VGG-19	0.907	0.924	0.790	0.969	**0.844**	0.915	**0.968**
1	AlexNet	0.925	0.830	0.632	0.972	0.751	0.878	0.943
	DenseNet-121	0.889	0.725	0.505	0.953	0.644	0.807	0.888
	ResNet-50	0.944	0.883	0.718	0.980	**0.816**	0.913	0.963
	ResNext-50	0.944	0.842	0.653	0.979	0.772	0.893	0.960
	ChexNet	0.944	0.853	0.671	0.979	0.784	0.899	0.966
	EfficientNet-b3	0.925	0.818	0.617	0.972	0.740	0.872	0.962
	VGG-19	0.907	0.865	0.680	0.967	0.777	0.886	0.960

Figure 6. ROC Curves of Designed Models

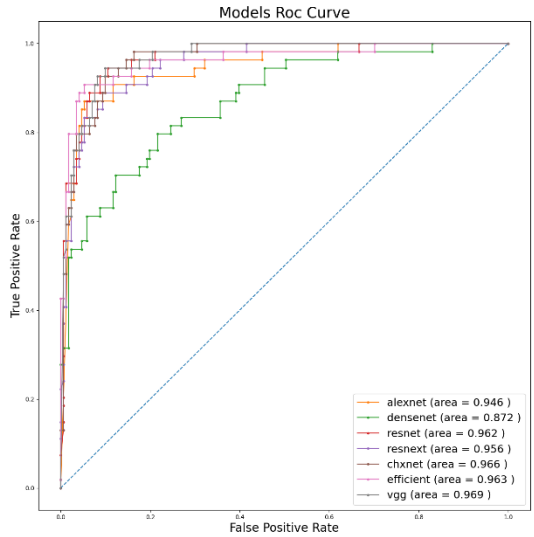
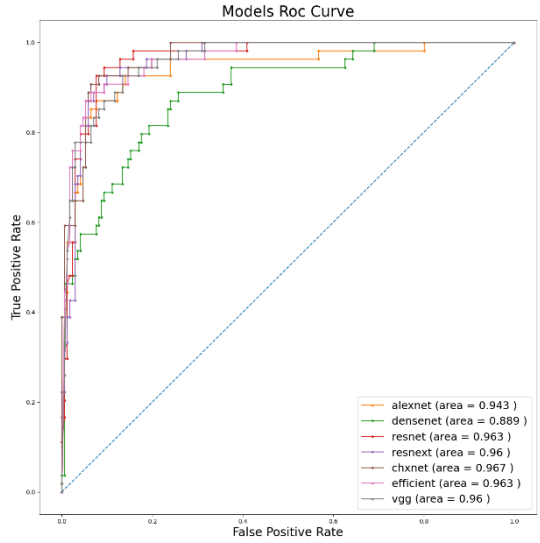
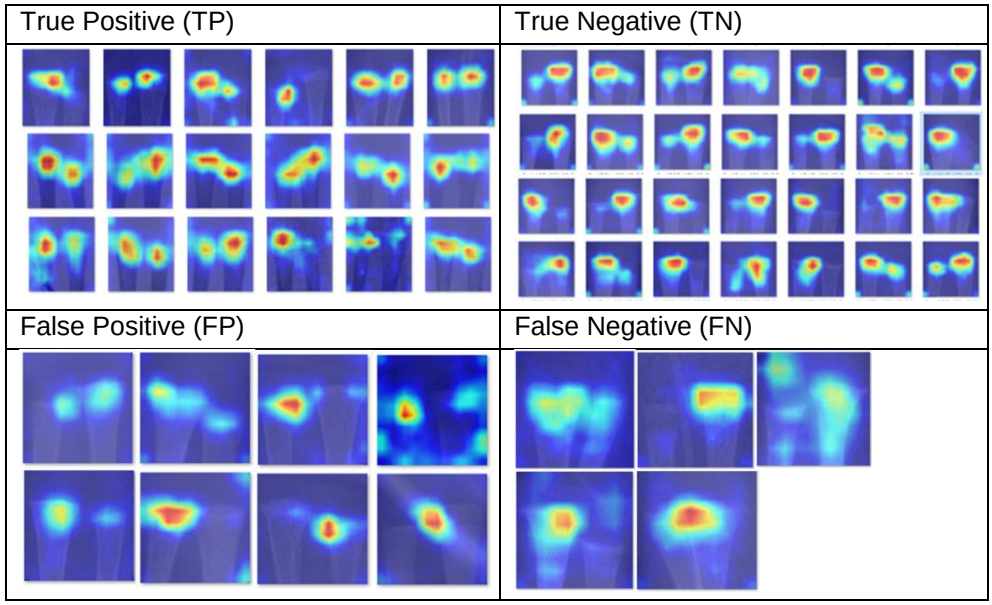
ROI 0	
ROI 1	

Figure 7. Gradcam Images



REFERENCES

1. Rustico, S.E., A.C. Calabria, and S.J. Garber, *Metabolic bone disease of prematurity.* J Clin Transl Endocrinol, 2014. 1(3): p. 85-91.
2. Chacham, S., et al., *Metabolic Bone Disease in Premature Neonates: An Unmet Challenge.* J Clin Res Pediatr Endocrinol, 2020. 12(4): p. 332-339.
3. Faienza, M.F., et al., *Metabolic Bone Disease of Prematurity: Diagnosis and Management.* Front Pediatr, 2019. 7: p. 143.
4. Kavurt, S., et al., *Evaluation of radiologic evidence of metabolic bone disease in very low birth weight infants at fourth week of life.* J Perinatol, 2021. 41(11): p. 2668-2673.
5. Chinoy, A., M.Z. Mughal, and R. Padidela, *Metabolic bone disease of prematurity: causes, recognition, prevention, treatment and long-term consequences.* Arch Dis Child Fetal Neonatal Ed, 2019. 104(5): p. F560-f566.
6. Stewart, K., et al., *Screening for Metabolic Bone Disease in Preterm Infants.* ICAN: Infant, Child, & Adolescent Nutrition, 2015. 7(5): p. 229-232.
7. Moreira, A., et al., *Metabolic Bone Disease of Prematurity.* NeoReviews, 2015. 16(11): p. e631-e641.
8. Rayannavar, A. and A.C. Calabria, *Screening for Metabolic Bone Disease of prematurity.* Semin Fetal Neonatal Med, 2020. 25(1): p. 101086.
9. Chang, C.Y., et al., *Imaging Findings of Metabolic Bone Disease.* Radiographics, 2016. 36(6): p. 1871-1887.
10. Koo, W.W., et al., *Skeletal changes in preterm infants.* Arch Dis Child, 1982. 57(6): p. 447-52.
11. Kahn, C.E., Jr., *From Images to Actions: Opportunities for Artificial Intelligence in Radiology.* Radiology, 2017. 285(3): p. 719-720.
12. Chartrand, G., et al., *Deep Learning: A Primer for Radiologists.* Radiographics, 2017. 37(7): p. 2113-2131.
13. Willemink, M.J., et al., *Preparing Medical Imaging Data for Machine Learning.* Radiology, 2020. 295(1): p. 4-15.
14. Erickson, B.J., et al., *Machine Learning for Medical Imaging.* Radiographics, 2017. 37(2): p. 505-515.
15. Moore, M.M., et al., *Machine learning concepts, concerns and opportunities for a pediatric radiologist.* Pediatr Radiol, 2019. 49(4): p. 509-516.
16. Rosendahl, K., et al., *Revisiting the radiographic assessment of osteoporosis-Osteopenia in children 0-2 years of age. A systematic review.* PLoS One, 2020. 15(11): p. e0241635.
17. Krizhevsky, A., I. Sutskever, and G.E. Hinton, *ImageNet classification with deep convolutional neural networks*, in *Proceedings of the 25th*

International Conference on Neural Information Processing Systems - Volume 1. 2012, Curran Associates Inc.: Lake Tahoe, Nevada. p. 1097-1105.
18. Huang, G., et al. *Densely Connected Convolutional Networks*. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.
19. He, K., et al. *Deep Residual Learning for Image Recognition*. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.
20. Xie, S., et al. *Aggregated Residual Transformations for Deep Neural Networks*. in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.
21. Rajpurkar, P., et al., *CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning.* 2017.
22. Simonyan, K. and A. Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition.* CoRR, 2015. abs/1409.1556.
23. Tan, M. and Q.V. Le, *EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks.* ArXiv, 2019. abs/1905.11946.
24. Paszke, A., et al. *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. in *NeurIPS*. 2019.
25. Handelman, G.S., et al., *Peering Into the Black Box of Artificial Intelligence: Evaluation Metrics of Machine Learning Methods.* AJR Am J Roentgenol, 2019. 212(1): p. 38-43.
26. Gaio, P., et al., *Incidence of metabolic bone disease in preterm infants of birth weight <1250 g and in those suffering from bronchopulmonary dysplasia.* Clin Nutr ESPEN, 2018. 23: p. 234-239.
27. Avila-Alvarez, A., et al., *Metabolic Bone Disease of Prematurity: Risk Factors and Associated Short-Term Outcomes.* Nutrients, 2020. 12(12).
28. Oestreich, A.E., *Concave distal end of ulna metaphysis alone is not a sign of rickets.* Pediatr Radiol, 2015. 45(7): p. 998-1000.
29. Meda, K.C., S.S. Milla, and B.S. Rostad, *Artificial intelligence research within reach: an object detection model to identify rickets on pediatric wrist radiographs.* Pediatr Radiol, 2021. 51(5): p. 782-791.
30. You, S.K., et al., *Metabolic bone disease in preterm infants: Relationship between radiologic grading in the wrist and serum biochemical markers.* Diagn Interv Imaging, 2017. 98(11): p. 785-791.

국문초록

서론: 미숙아 대사성 골질환은 미숙아가 겪는 중요한 합병증 중 하나로 정확한 진단 및 적절한 시점에서의 치료적 개입이 필요한 질환이다.

목적: 본 연구는 미숙아 대사성 골질환의 진단을 용이하게 하고자 손목 x-ray 영상 정보를 바탕으로 미숙아 대사성 골질환 진단 딥러닝 모델을 구축하고자 한다.

방법: 2010년부터 2020년 사이에 서울대학교 어린이병원에서 1500g 미만으로 출생한 미숙아들 중 신생아중환자실에 입실한 환자들을 대상으로 연구가 진행되었다. 인구학적 정보, 임상 정보, 생후 4-8주 사이에 촬영된 손목 x-ray 영상들은후향적으로 수집되었다. 딥러닝 모델 학습을 위해 두 가지 관심 영역 ('ROI 0'과 'ROI 1')의 어노테이션이 완료되었다. 임상정보는 미숙아 대사성 골질환과 연관된 인자들을 분석하고자 사용되었고, 이를 통해 연구 모집단의 대표성을 확인하고자 하였다. 수집된 손목 x-ray 영상은 딥러닝을 통한 진단 프로그램을 개발하기 위한 학습데이터로 사용되었다. 프로그램 개발을 위해 AlexNet, DenseNet-121, ResNet-50, ResNext-50, VGG-19, CheXNet, EfficientNet-b3 딥러닝 architecture 가 사용되었다.

결과: 모집단 중 14.3% (116/814)가 생후 4-8주 사이에 미숙아 대사성 골질환으로 진단되었다. 생후 4-8주 이내에 한 번이라도 손목 영상에서 대사성 골질환으로 진단된 경우와 그렇지 않은 경우를 두 군으로 비교하였고, 출생체중 1000g 미만 (82.8% vs. 37.5%, p=0.000), 재태주수 28주 미만 (75.0% vs. 29.5%, p=0.000), 정맥영양 공급 기간 28일 이상 (49.1% vs, 12.0%, p=0.000)이 질환을 겪은 군에서 유의미하게 높은 빈도임이 확인되어, 대사성 골질환의 위험인자로 확인되었다. 이는 이미 잘 알려진 미숙아 대사성 골질환의 위험인자와 일치하며, 이를 통해 모집단이 일반적인 미숙아 집단을 대표할 수 있음을 확인하였다. 더불어 학습에 사용된 ground truth의 신뢰도 또한 입증할 수 있었다. 'ROI 0' 을 이용하여 EfficientNet-b3와 VGG-19를 통해 개발한 진단 모델이 가장 뛰어난 성능을 나타내며, 최대값의 F1 스코어 (0.844)와 AUROC 값 (EfficientNet-b3: 0.962, VGG-19: 0.968)을 보였다., 두 모델의 민감도는 0.907, 특이도는 0.924, 양성 예측도는 0.790, 음성 예측도는 0.969, 정확도는 0.915였다.

결론: 본 연구를 통해 미숙아 대사성 골질환 진단을 위한 딥러닝 모델이 개발되었고 민감도는 0.907, 특이도는 0.924, 정확도는 0.915이다. 향후에 이러한 진단기법이 실제 임상에 적용된다면, 특히나 임상경력이 적

은 임상의의 경우에도 질환의 진단이 정확하고 간편하게 이루어질 수 있을 것으로 생각하며, 이를 통해 치료 및 예방을 위한 적절한 개입이 가능해질 것으로 기대한다.

.....

주요어: 미숙아 대사성 골 질환, 인공지능, 딥러닝, 미숙아, 손목 x-선

학번: 2020-29040