mathematics

MDPI

Article

A Novel Method for Medical Predictive Models in Small Data Using Out-of-Distribution Data and Transfer Learning

Inyong Jeong [1], Yeongmin Kim [1], Nam-Jun Cho [2], Hyo-Wook Gil [2] and Hwamin Lee [1,*]

[1] Department of Biomedical Informatics, Korea University College of Medicine, Seoul 02708, Republic of Korea; 5454dls@korea.ac.kr (I.J.); ming8387@korea.ac.kr (Y.K.)
[2] Department of Internal Medicine, Soonchunhyang University Cheonan Hospital, Cheonan 31151, Republic of Korea; hwgil@schmc.ac.kr (H.-W.G.)
* Correspondence: hwamin@korea.ac.kr

Abstract: Applying deep learning to medical research with limited data is challenging. This study focuses on addressing this difficulty through a case study, predicting acute respiratory failure (ARF) in patients with acute pesticide poisoning. Commonly, out-of-distribution (OOD) data are overlooked during model training in the medical field. Our approach integrates OOD data and transfer learning (TL) to enhance model performance with limited data. We fine-tuned a pre-trained multi-layer perceptron model using OOD data, outperforming baseline models. Shapley additive explanation (SHAP) values were employed for model interpretation, revealing the key factors associated with ARF. Our study is pioneering in applying OOD and TL techniques to electronic health records to achieve better model performance in scenarios with limited data. Our research highlights the potential benefits of using OOD data for initializing weights and demonstrates that TL can significantly improve model performance, even in medical data with limited samples. Our findings emphasize the significance of utilizing context-specific information in TL to achieve better results. Our work has practical implications for addressing challenges in rare diseases and other scenarios with limited data, thereby contributing to the development of machine-learning techniques within the medical field, especially regarding health inequities.

Keywords: transfer learning; out-of-distribution data; machine learning; limited medical data; acute respiratory failure; acute pesticide poisoning

MSC: 60E99; 62B86; 68T99; 92C50

Citation: Jeong, I.; Kim, Y.; Cho, N.-J.; Gil, H.-W.; Lee, H. A Novel Method for Medical Predictive Models in Small Data Using Out-of-Distribution Data and Transfer Learning. *Mathematics* **2024**, *12*, 237. https://doi.org/10.3390/math12020237

Academic Editors: Ivan Izonin, Addisson Salazar, Stephane Chretien, Leszek Rutkowski and Faheim Sufi

Received: 15 December 2023
Revised: 1 January 2024
Accepted: 9 January 2024
Published: 11 January 2024

1. Introduction

Machine learning has emerged as a prominent field in the current medical research landscape. However, constructing effective machine-learning models, particularly with deep-learning (DL) techniques, often requires large amounts of data [1]. Acquiring the necessary volume of datasets can be time consuming and resource intensive, involving significant resource costs, including financial expenses. This challenge is especially pronounced in specialized medical fields, where data acquisition may be hindered by low-prevalence diseases, regional inequalities or standardization challenges [2]. When the ratio of training samples to the Vapnik–Chervonenkis (VC) dimensions of a learning machine is less than 20, it is considered a small sample size [3–6]. VC dimensions measure the capacity of a classifier, representing the cardinality of the greatest collection of points, which the procedure can break [4]. However, these theories do not apply to real-world scenarios with limited datasets for machine-learning models, as they primarily focus on generic machine learning with a high number of training samples [7]. Using limited datasets risks inadequate model training, potentially reducing the likelihood of achieving global minima [8]. Additionally, random weight initialization in machine-learning models may introduce further uncertainty, thereby highlighting the necessity for meticulous weight

initialization in scenarios with limited data [9–12]. Moreover, small datasets pose various challenges, including overfitting problems, the significant impact of noise components, missing values, outliers and sharp fluctuations in variables within the dataset, resulting in low generalization ability [13].

Despite these challenges, small datasets possess intrinsic value, and ongoing research endeavors are addressing the mentioned issues [14,15]. Traditionally, augmentation methods, such as changing direction or adjusting angles, have been prevalent in the image domain [16]. Additionally, oversampling strategies are commonly employed in tabular data to increase the number of patient data, especially for minority groups [17]. Recent approaches utilizing generative adversarial networks or diffusion models for data synthesis aim to overcome these issues [18–20]. However, these strategies have yet to resolve trust issues related to generated data and require substantial resources. They are also predominantly limited to the field of images [21,22]. Other various algorithms, such as ensemble learning and input-doubling method [13,23–26], have also been actively researched. However, they mostly face similar challenges, and there is no clear consensus on technically feasible solutions, necessitating further research [13,27,28].

Furthermore, prior knowledge has improved predictive accuracy over random weight initialization in data-deficient contexts. As a result, the medical field has been actively exploring transfer learning (TL) to achieve this [29]. TL involves refining a pre-trained model on extensive datasets by adapting insights from one related task to another. TL can produce robust models capable of operating with sparse data, shortening training durations and improving model generalization [30]. However, its primary medical sector use is limited to imaging tasks, while electronic health record (EHR) data are often structured in tabular form, making it challenging to acquire compatible large-scale datasets for TL applications [29–32].

Meanwhile, the medical field frequently encounters diverse instances of out-of-distribution (OOD) data [33]. OOD data are generated from a distribution, which deviates from the one on which the model was initially trained [34]. In medical practice, OOD data are common in various scenarios, e.g., when data from different hospitals are used for external validation, when training and evaluation data are segregated, or when integrating data reflective of varying patient conditions or environments, even within the same medical condition [35]. OOD data often follow a distinct statistical distribution compared to in-distribution data and may exhibit contextual disparities [34]. The utilization of OOD data has enhanced machine models' overall performance and robustness when employed in the right context [9,10,36].

This study presents an innovative method, which uses OOD data and TL to overcome data scarcity in machine-learning model development. We propose a machine-learning model for predicting acute respiratory failure (ARF) in patients with acute pesticide poisoning, incorporating a unique model development approach. Acute pesticide poisoning is a public health concern worldwide, and it is often accompanied by fatal outcomes [37]. ARF is a major cause of mortality in patients with acute pesticide poisoning, and it is known that the clinical course differs based on the category of pesticide, ingestion amount and underlying disease [38]. Since acute pesticide poisoning is rare, a lack of clinical experience can prevent predicting the prognosis of patients. Therefore, an ARF prediction model is essential for the timely treatment of patients with acute pesticide poisoning [39,40]. Acute pesticide poisoning is more prevalent in rural areas than in urban areas, leading to variations in data based on the location of medical institutions. Collecting data is exceptionally challenging due to the infrequency of cases, making our novel approach well suited for this problem [30].

The main contributions of this paper can be summarized as follows:

1. Introducing a highly intuitive and simple idea and assessing the potential utility of OOD data in creating pre-trained models for TL.
2. Experimentally validating the effectiveness of OOD and TL in small medical datasets while minimizing artificial data manipulations, such as data generation.

3. Developing a predictive model for ARF in patients with acute pesticide poisoning using the proposed method, showcasing low bias and high performance.

The remainder of the paper is organized as follows. Section 2 encompasses the study population, labeling, feature selection, handling of outliers and missing values, and modeling. Section 3 presents the study participants' characteristics, model performance and model interpretation. Section 4 engages in a thorough analysis of the outcomes, addressing limitations and future research. Lastly, Section 5 provides a summary of the study's contributions and implications for the field. Additionally, in Appendix A, we present abbreviation descriptions (Table A1) along with tables and figures, which may aid in the understanding of the paper. In Appendix B, we offer additional experiments, which support and reinforce the experiments conducted in the manuscripts.

2. Materials and Methods

2.1. Study Population

The study was conducted on 129,953 patients aged 19 years and older admitted to the general ward at Korea University Anam Hospital between January 2015 and December 2021. To distinguish patients who experienced ARF from those who did not, we excluded 1508 patients who experienced ARF but had unclear onset times. These patients had not ingested pesticides and were considered as OOD data collected from different regions or hospitals.

A retrospective observational cohort study was conducted on 1081 patients with acute pesticide poisoning who were admitted to Soonchunhyang University Cheonan Hospital between January 2015 and December 2020. To ensure reliable results, exclusion criteria were established based on previous studies [37,41]. First, patients under the age of 19 were excluded, as were those who had been poisoned by paraquat-based pesticides, which are known to cause ARF within a short time. Considering the pattern of ARF occurrence and the study design, patients who were diagnosed with ARF within 1 h of admission or 72 h after admission were also excluded, as were patients with a "Do Not Resuscitate" status due to mechanical ventilator refusal (Figure 1). The final study cohort included 803 patients with acute pesticide poisoning.

The Institutional Review Boards (IRB) of Korea University Anam Hospital (IRB number: 2023AN0145) and Soonchunhyang University Cheonan Hospital (IRB number: 2020-02-016) reviewed and approved this study. The study was conducted following the principles outlined in the Helsinki Declaration.

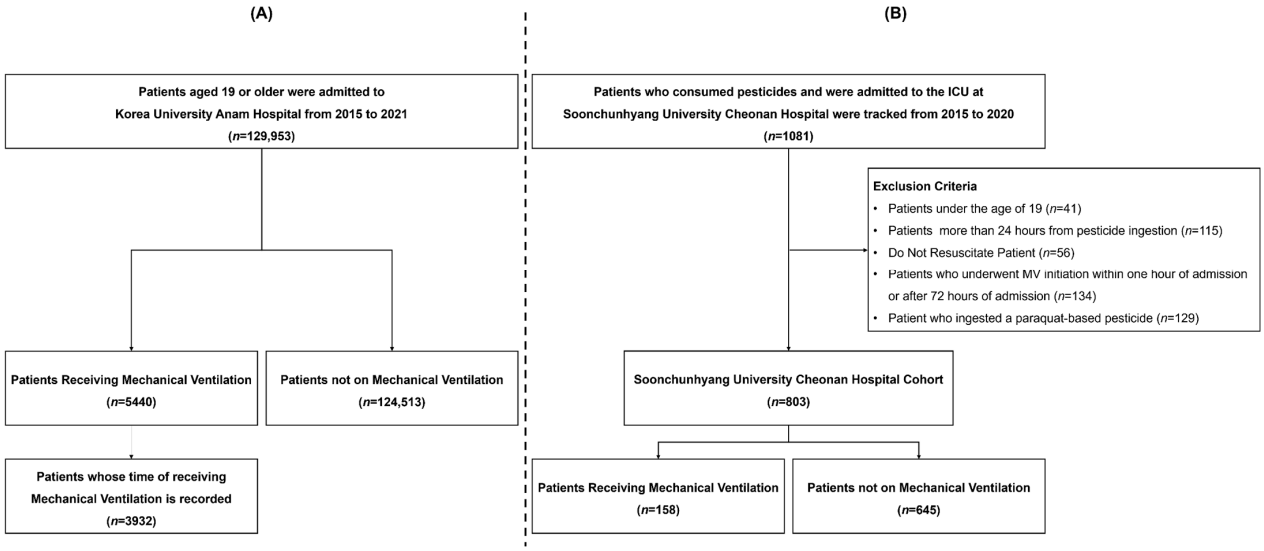

Figure 1. Study flowchart. (**A**) Korea University Anam Hospital, (**B**) Soonchunhyang University Cheonan Hospital.

2.2. Labeling

We considered the time of receiving mechanical ventilation as the onset of acute ARF. The study utilized data from two hospitals, each exhibiting distinct characteristics. As a result, specific research designs were implemented tailored to each dataset. Korea University Hospital data showed a notable scarcity of ARF cases, leading to a data imbalance issue.

The specific approach adopted to address this issue is illustrated in Figure 2. Patients who experienced ARF were labeled "1", with a prediction time of 1–72 h before the onset of the condition. Patients who did not experience ARF were labeled "0", with a prediction timeframe of 143–72 h before discharge to mitigate the uncertainty associated with the potential later onset after discharge. Data points outside the defined prediction timeframe were excluded, effectively rectifying the data imbalance issue.

Conversely, Soonchunhyang University Cheonan Hospital patients exhibited a different pattern, with ARF cases being more prevalent within 72 h of admission. This resulted in a less severe data imbalance. To maintain rigorous evaluation criteria and account for this pattern, a prediction timeframe of 1–72 h after admission was applied. Patients who experienced ARF were labeled "1", while those who did not experience this condition were labeled "0".

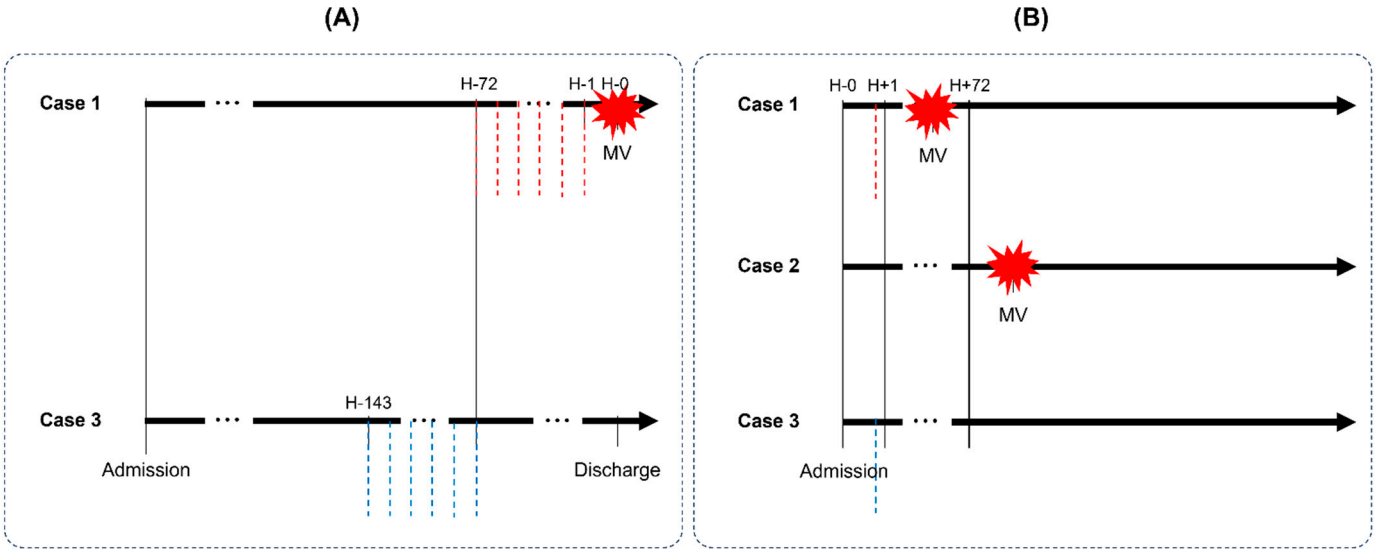

Figure 2. Study design for each hospital. "MV" refers to mechanical ventilation. Cases 1 and 2 pertain to patients receiving mechanical ventilation, with Case 2 subsequently excluded based on exclusion criteria. Case 3 corresponds to patients not on mechanical ventilation. (**A**,**B**) represent Korea University Anam Hospital and Soonchunhyang University Cheonan Hospital, respectively. The red and blue dashed lines indicate the prediction time points labeled as "1" and "0", respectively. The red star-shaped symbols signify the occurrence of mechanical ventilation.

2.3. Feature Selection

The feature selection process was informed by prior studies [15,19]. We additionally consulted with experts to determine relevant features. We then excluded features, which were not commonly applicable to TL. For example, the Glasgow Coma Scale (GCS), considered a critical feature in past research, was excluded from our study due to its high rate of missing data in the Korea University Anam Hospital dataset. Considering the limited sample size and the need for rapid data assessment, we prioritized features, which are typically measured in the majority of patients within an hour, ensuring minimal missing data. As a result, we only selected variables missing from <5% of the Soonchunhyang University Cheonan Hospital data. The selected features included age, sex, systolic blood pressure (SBP), diastolic blood pressure (DBP), respiratory rate, body temperature, serum creatinine, hemoglobin, total carbon dioxide (Total CO_2), pH, pCO_2, pO_2, base excess (BE),

lactate, category of pesticide and amount of ingestion. In this context, "sex" refers to the sex assigned at birth.

2.4. Handling of Outliers and Missing Values

To tackle potential outliers, values falling below the 2.5th percentile or exceeding the 97.5th percentile for each attribute were considered outliers and treated as missing values to eliminate their potential influence on the whole dataset. Subsequently, the multiple imputation by chained equations (MICE) algorithm was used to impute the missing data. MICE is widely used to generate imputations, which closely resemble true distributions when the rate of missing values is low [42]. Following this, robust scaling was applied. Notably, MICE and robust scaling computations were exclusively performed on the training data throughout all phases of the learning process.

Before performing the above pre-processing, the data obtained from Korea University Anam Hospital contained numerous missing values, necessitating additional pre-processing. First, we organized the features daily. Systolic blood pressure (SBP), diastolic blood pressure (DBP), respiratory rate and body temperature were arranged daily using the highest recorded values. The remaining attributes were assigned the last recorded values. Despite these efforts, any remaining missing data were then imputed by referencing the most recent observations. Furthermore, the pesticide category and ingestion amount were not available in the Korea University Anam Hospital dataset and were uniformly set to zero.

2.5. Modeling and Performance Evaluation

Predicting ARF in patients with acute pesticide poisoning is a challenging task due to the limited availability of such cases. This study used a large-scale OOD dataset of patients without acute pesticide poisoning to overcome this challenge to enhance ARF prediction. The study employs various machine-learning models, including logistic regression (LR), random forest (RF), extreme gradient boosting (XGB), light gradient-boosting machine (LGBM) and a multi-layer perceptron (MLP). The regression analysis, ensemble method and neural network models considered in our study are among the most commonly utilized models in the development of clinical prediction models [43,44]. Furthermore, since the data we are working with are not in the form of time series or images, we did not consider models such as recurrent neural networks or convolutional neural networks. Our novel approach to TL using OOD data is illustrated in Figure 3.

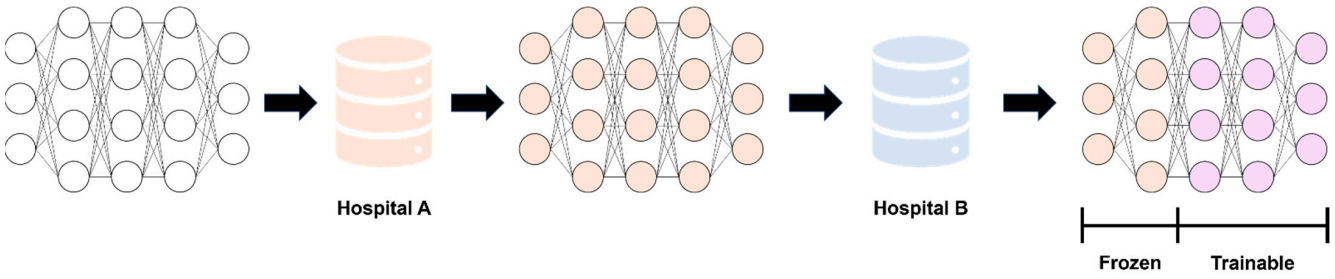

Figure 3. Transfer learning process. We used a model pre-trained on OOD data to initialize the initial weights and adjust the number of trainable layers. Red represents the data from Hospital A and the corresponding training results of the model using that data. Blue signifies the data from Hospital B and the model's training outcomes based on it. During the learning process, layers frozen during training by Hospital B maintain their red color, indicating that training by Hospital B did not influence them. In contrast, layers that underwent training exhibit a mix of red and blue, resulting in a purple hue.

The first step of the approach is to develop an MLP model, which can predict ARF in patients without acute pesticide poisoning. This MLP model is a pre-trained model, serving as a foundation for fine-tuning using data from patients with acute pesticide poisoning.

During the fine-tuning process, the number of trainable layers is adjusted, and each model variant is evaluated systematically. The initial MLP model consists of five dense layers, including the output layer. The number of trainable dense layers is varied systematically to create models ranging from TL1 to TL5. The specific model architecture can be examined in Figure A3.

To ensure the reliability of the models, cross-validation is carried out due to the limited number of patients in the cohort. The dataset is divided into five groups, with the ratio of patients with ARF to patients without ARF maintained in each group. Group 5 underwent early stopping in DL, while the remaining groups (Groups 1–4) were used for four-fold cross-validation, as shown in Figure A4.

We considered key performance metrics during model evaluation, such as the area under the receiver operating characteristic (AUROC) and the F1 score. For the final evaluation, a comprehensive range of performance metrics is considered for the best performing model, which includes accuracy, precision, recall, F1 score, negative predictive value (NPV), Matthews correlation coefficient (MCC), AUROC and the area under the precision–recall curve (AUPRC). In addition to these quantitative metrics, a visual assessment is conducted to comprehensively understand the model's performance. This visual inspection involves the examination of confusion matrices, AUROC curves and AUPRC curves. It provides valuable insights into the model's performance, behavior and strengths.

2.6. Statistical Analysis and Model Interpretation

The basic statistics of the datasets from both hospitals were thoroughly examined. A *t*-test was conducted at a significance level of 0.05 to determine potential between-hospital differences. Each hospital's cases were designated as "1" (indicating ARF) and "0" (indicating non-ARF) and examined separately. Subsequent *t*-tests were performed at a significance level of 0.05 on data subsets to determine the significance of the observed differences.

To better understand how the model works and which features are most important, we identified the features, which received high evaluations from the model. We also used Shapley additive explanation (SHAP) values to confirm the clinical significance of the model's learning process. These SHAP values help us understand how each feature contributes to the model's predictions, making it easier to interpret model-guided decisions. This step is important in determining the practical relevance of the model's findings in a clinical setting.

3. Results

3.1. Study Participants' Characteristics

After addressing the outliers, we provide in Table A2 the missing values for each feature and their corresponding proportions. Our study approach comprised selecting features with clinically significant relevance while ensuring that the proportion of missing values did not exceed 5%. Soonchunhyang University Cheonan Hospital initially selected features with less than 5% missing data. However, additional missing values were introduced after handling the outliers, causing some features to exceed the 5% threshold. Table A2 includes the GCS—which was otherwise excluded from feature selection—for comparative purposes.

Table 1 includes each feature's mean and standard deviation. A statistical analysis was conducted using *p*-values obtained from the *t*-tests to verify between-hospital differences. The *t*-test results indicated significant differences for all features, suggesting the two datasets were OOD. For the sex feature, the table presents the number and percentage of males. A chi-squared test confirmed the differences detailed in the table. This table offers a comprehensive overview of the statistical differences between the datasets, emphasizing the OOD relationship and the specific comparison for the sex feature.

Table 1. Mean and standard deviation according to the feature. "*" means statistically significant under a significance level of 0.05.

Feature	Korea University Anam Hospital (n = 12,059)		Soonchunhyang University Cheonan Hospital (n = 803)		p-Value
	Mean	SD	Mean	SD	
Age, year	68.48	16.13	61.54	15.70	<0.0001 *
Sex (%) [1]	6762	56.07	500	62.27	0.0007 *
Systolic BP, mmHg [2]	123.38	16.26	133.90	23.86	<0.0001 *
Diastolic BP, mmHg	73.28	10.80	78.18	12.97	<0.0001 *
Respiratory rate, bpm	19.36	2.72	19.36	1.83	<0.0001 *
Heart rate, bpm	36.92	0.45	36.42	0.56	<0.0001 *
Serum Cr, mg/dL [3]	1.12	0.83	0.86	0.27	<0.0001 *
Hemoglobin, g/dL	10.79	1.97	14.04	1.65	<0.0001 *
Total CO_2, mmol/L	23.45	3.97	22.26	3.39	<0.0001 *
Arterial pH	7.43	0.04	7.38	0.06	<0.0001 *
pCO_2, mmHg	32.07	5.80	37.11	5.72	<0.0001 *
pO_2, mmHg	93.02	27.65	85.66	18.11	<0.0001 *
BE, mmol/L [4]	−1.51	3.28	−2.33	4.08	<0.0001 *
Lactate, mmol/L	1.98	1.18	2.88	1.90	<0.0001 *

[1] Sex, male; [2] BP, blood pressure; [3] Cr, creatinine; [4] BE, base excess.

Tables 2 and 3 present key statistics, such as means, standard deviations and statistical significance, highlighting the differences between patients with ARF and patients without ARF in both hospital datasets. For the Korea University Anam Hospital dataset, data pre-processing intentionally introduced distinctions between patients with ARF and patients without ARF, resulting in significant disparities in all features. However, in the Soonchunhyang University Cheonan Hospital dataset, some features were not significantly different between the patients with ARF and patients without ARF.

Table 2. Differences between patients with ARF and patients without ARF at Korea University Anam Hospital. "*" means statistically significant under a significance level of 0.05.

Feature	Patients without ARF (n = 645)		Patients with ARF (n = 158)		p-Value
	Mean/N	SD/%	Mean/N	SD/%	
Age, year	68.66	16.42	67.59	14.53	<0.0001 *
Sex (%) [1]	54.05	54.61	1357	62.80	<0.0001 *
Systolic BP, mmHg [2]	123.51	15.89	122.67	18.14	<0.0001 *
Diastolic BP, mmHg	73.48	10.61	72.21	11.69	<0.0001 *
Respiratory rate, bpm	19.15	2.40	20.51	3.86	<0.0001 *
Heart rate, bpm	36.92	0.44	36.94	0.53	<0.0001 *
Serum Cr, mg/dL [3]	1.07	0.77	1.41	1.03	<0.0001 *
Hemoglobin, g/dL	10.87	1.95	10.37	2.03	<0.0001 *
Total CO_2, mmol/L	23.58	3.90	22.75	4.26	<0.0001 *
Arterial pH	7.43	0.04	7.43	0.05	<0.0001 *
pCO_2, mmHg	32.03	5.69	32.31	6.40	<0.0001 *
pO_2, mmHg	92.92	27.09	93.54	30.41	<0.0001 *
BE, mmol/L [4]	−1.46	3.20	−1.75	3.70	<0.0001 *
Lactate, mmol/L	1.92	1.12	2.26	1.42	<0.0001 *

[1] Sex, male; [2] BP, blood pressure; [3] Cr, creatinine; [4] BE, base excess.

Table 3. Differences between patients with ARF and patients without ARF at Soonchunhyang University Cheonan Hospital. "*" means statistically significant under a significance level of 0.05.

Feature	Patients without ARF (n = 645)		Patients with ARF (n = 158)		p-Value
	Mean/N	SD/%	Mean/N	SD/%	
Age, year	59.91	15.81	68.18	13.35	<0.0001 *
Sex (%) [1]	408	63.26	92	58.23	0.2815
Systolic BP, mmHg [2]	133.63	23.25	135.00	26.32	<0.5583
Diastolic BP, mmHg	78.55	12.67	76.63	14.06	<0.1257
Respiratory rate, bpm	19.36	1.75	19.34	2.19	0.9141
Heart rate, bpm	36.45	0.52	36.28	0.67	<0.0047 *
Serum Cr, mg/dL [3]	0.83	0.27	0.97	0.27	<0.0001 *
Hemoglobin, g/dL	14.09	1.63	13.83	1.70	0.0988
Total CO_2, mmol/L	22.69	3.19	20.51	3.61	<0.0001 *
Arterial pH	7.39	0.06	7.36	0.08	<0.0001 *
pCO_2, mmHg	37.20	5.71	36.68	5.76	0.3319
pO_2, mmHg	85.37	16.97	87.01	22.65	0.4274
BE, mmol/L [4]	−1.85	3.88	−4.38	4.28	<0.0001 *
Lactate, mmol/L	2.84	1.82	3.03	2.24	0.3526

[1] Sex, male; [2] BP, blood pressure; [3] Cr, creatinine; [4] BE, base excess.

Table A3 provides an insightful overview of the distribution of pesticide-related features in the Soonchunhyang University Cheonan Hospital dataset, offering a better understanding of the features' characteristics. These tables comprehensively illustrate the differences between patients with ARF and patients without ARF and provide insights into the distribution of pesticide-related features in the dataset.

3.2. Model Performance

Each model's major performance metrics, including their AUROC and F1 scores, are included in Table 4. The models have generally high AUROC values, but the wide range of confidence intervals raises concerns about the reliability of their performance. This variability in performance can be attributed to differences between the training and test sets, especially in limited datasets. The RF model has the highest AUC among the traditional models, while the LR model has the narrowest confidence interval. The MLP model has the lowest mean and the widest confidence interval.

Table 4. Model performance. In transfer learning, the term "numbers" refers to the number of trainable dense layers.

Model	AUROC		F1	
	Mean	95% CI	Mean	95% CI
LR	0.8665	0.8384–0.8946	0.4936	0.3565–0.6308
RF	0.8767	0.8189–0.9346	0.4173	0.2567–0.5780
XGB	0.8620	0.8238–0.9003	0.5179	0.4039–0.6319
LGBM	0.8627	0.8101–0.9153	0.5511	0.4420–0.6602
MLP	0.8361	0.7282–0.9439	0.4411	0.1924–0.6898

Table 4. *Cont.*

Model	AUROC		F1	
	Mean	95% CI	Mean	95% CI
TL				
5	0.9023	0.8760–0.9286	0.5539	0.4413–0.6665
4	0.8884	0.8513–0.9255	0.5672	0.4127–0.7216
3	0.8679	0.8344–0.9014	0.5435	0.4477–0.6392
2	0.8654	0.8107–0.9201	0.5513	0.5152–0.5873
1	0.8562	0.8008–0.9115	0.5228	0.4362–0.6094

Notably, the TL model, which has the same structures as the MLP model, significantly outperforms the existing models. The TL approach remarkably narrows the confidence interval, substantially enhancing overall model performance. Figure 4 visually illustrates the comparative performance of each model. Table A4 illustrates the performance when GCS is included as a feature. Incorporating GCS as a feature significantly enhances the performance across all models. This confirms the crucial importance of GCS as a feature in patients with acute pesticide poisoning.

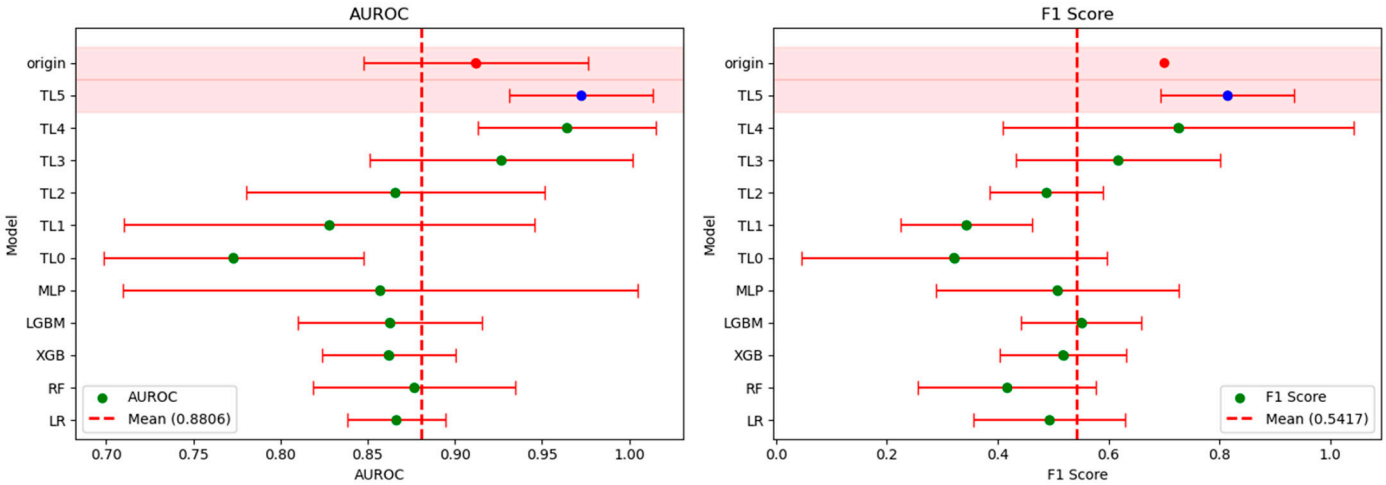

Figure 4. Presentation of AUROC and F1 scores for each model with 95% confidence intervals. Red dots denote the performance of the model reported in prior studies, while blue dots represent the performance of the best model derived from new proposed approach.

Table 5 provides detailed performance metrics for the MLP model, which has the same structure as the high-performing TL5 model and uses Group 4 evaluation data. These metrics offer a more detailed view of the model's performance and effectiveness in the specific evaluation context. The table includes a comprehensive set of performance metrics, such as accuracy, precision, recall, F1 score, NPV, MCC, AUROC and AUPRC. Figure A1 visually compares the models' performance, highlighting the confusion matrix, AUROC curve and AUPRC curves. Notably, there is a significant improvement in precision and recall for the TL5 model, highlighting substantial enhancements in its overall performance.

Table 5. Model performance for Group 4.

Model	Accuracy	Precision	Recall	F1	NPV	MCC	AUROC	AUPRC
MLP	0.83	0.83	0.16	0.26	0.83	0.31	0.77	0.57
TL5	0.87	0.87	0.41	0.55	0.87	0.54	0.91	0.79

3.3. Model Interpretation

Figure A2 displays cases where there is a probability difference of 0.1 or more between the MLP and TL5 models. The red area represents patients with ARF, while the blue area represents patients without ARF. Therefore, if the model's performance is high, the probability should be higher in the red area and lower in the blue area. The observed trend suggests that, except for some cases, the probability increases for patients with ARF and decreases for patients without ARF. Ultimately, the model predicts with greater confidence that patients who experience ARF are more likely to do so, and patients who do not experience ARF are more likely to remain free from it. This indicates an increased discriminative ability of the model.

Figure 5 displays the model's SHAP values, highlighting the significant factors contributing to the development of ARF. The analysis provides the following insights:

1. High Cr, low TCO_2 and low DBP significantly contributed to the development of ARF.
2. Older age, low BE, high pCO_2 and high SBP may contribute to the development of ARF.
3. Glufosinate and organophosphates were more likely to contribute to the development of ARF than other pesticides.
4. Ingesting less than 100 cc carried a lower likelihood of developing ARF, while those who ingested 100–200 cc showed higher likelihood.

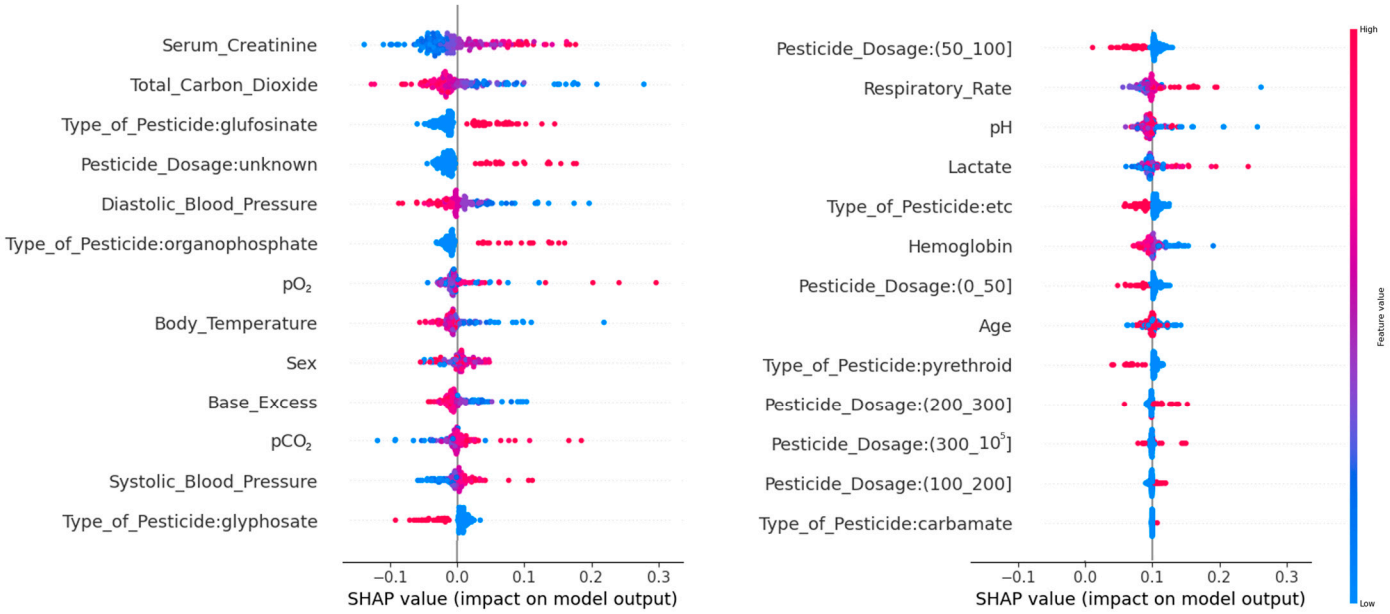

Figure 5. SHAP values for the TL5 model.

4. Discussion

Our study is a pioneering effort by the authors to apply OOD and TL techniques, commonly used in the image domain, to EHR, aiming to improve the performance of models with small sample sizes. Specifically, the authors developed a model for predicting ARF in patients with acute pesticide poisoning with minimized bias [37,41]. In cases of acute pesticide poisoning, MLP models face limitations due to insufficient data, resulting in low performance and wide confidence intervals [45]. In contrast, the newly proposed approach outperforms the MLP model and exhibits narrower confidence intervals. Additionally, the highest performance was achieved when the number of trainable layers was maximized. Maximizing the utilization of information tailored to the intended purpose is more advantageous than simply using information from a pre-trained model. The study also confirms the potential benefits of initializing weights using OOD data, particularly in cases of limited

data, instead of commonly used initialization [9–12]. Moreover, transfer learning showed its ability to enhance performance, even when data are scarce, as illustrated in Figure A1.

Simply examining retrospective data does not allow for a discussion on the mechanisms of ARF between patients with acute pesticide poisoning and those without. However, leveraging OOD data, the model may have learned more generalized and rough patterns regarding the deteriorating respiratory condition of patients. The weights configured in this manner are expected to be effective in facilitating TL effectively. To assess the importance of features, SHAP values were employed. Most of the results were consistent with the trends of feature importance identified in previous research. By combining the importance of individual features as indicated by SHAP values with factors such as pesticide category and ingestion amounts, future research can contribute to a better understanding of the mechanisms underlying ARF resulting from acute pesticide poisoning.

However, this study has some limitations. First, it is retrospective and based on data from a single institution. Future studies should address these limitations and expand the scope of data collection. Second, further study is needed to examine the best ways to use OOD data, investigate various TL application methods and develop strategies for handling differing features in different application contexts. For instance, there is a need for discussion on how to address challenges when important features, such as GCS in this study, are mostly missing, making them difficult to leverage in pre-training. Despite its limitations, this study contributes valuable new methodologies for managing limited data in studying rare diseases and comparable conditions, highlighting the significant promise of machine-learning techniques in advancing medical research.

5. Conclusions

This study pioneers the application of OOD and TL techniques in the EHR, particularly in scenarios characterized by limited data. We conducted this research with a focus on predicting acute respiratory failure in patients with acute pesticide poisoning. Our proposed approach surpasses conventional predictive models by leveraging OOD data in conjunction with pre-trained models, highlighting the substantial benefits of OOD data for weight initialization in settings where data are scarce. The outcomes of our experimentation suggest that our method holds promise as a viable alternative for effectively training models with limited data. When an appropriate OOD dataset is adeptly utilized, it introduces a compelling methodology for addressing data limitations in rare diseases and analogous scenarios. Future research should be expanded beyond these preliminary findings to refine transfer learning applications and formulate strategies for handling diverse data attributes across various medical scenarios. A key emphasis should be placed on addressing the challenge of managing crucial yet dissimilar features in prediction. In conclusion, our study makes a significant contribution by presenting innovative methodologies to navigate challenges posed by limited data in the study of rare diseases and similar conditions. We will conduct additional research to overcome the limitations discussed in this paper.

Author Contributions: Conceptualization, I.J.; Formal Analysis, I.J., Y.K., N.-J.C. and H.-W.G.; Investigation, I.J. and Y.K.; Methodology, I.J.; Writing, I.J., Y.K., N.-J.C., H.-W.G. and H.L.; Data Curation, N.-J.C. and H.-W.G.; Funding Acquisition, H.L.; Project Administration, H.L.; Resources, H.L.; Supervision, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the Basic Science Research Program of the National Research Foundation (NRF-2021R1A2C1009290), the ICAN (ICT Challenge and Advanced Network of HRD) program (IITP-2024-RS-2022-00156439), which is supervised by the IITP (Institute of Information and Communications Technology Planning and Evaluation), and a Korea University Grant (K2210721).

Data Availability Statement: The dataset used in this study is not publicly available. However, the data used in this study can be provided if there is a reasonable request made to the corresponding author. The code for generating the result of this study can be accessed at https://github.com/5454dls/OOD_TL (accessed on 1 January 2024). Furthermore, we have shared the results of a simple

validation of our approach using open datasets. These findings are also available in Appendix B and the same repository.

Acknowledgments: We would like to acknowledge the contributions of Minhyup Kim, a student at Korea University, in summarizing preliminary research.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

We compiled the figures and tables from our research results, which were not utilized in the main text, in Appendix A.

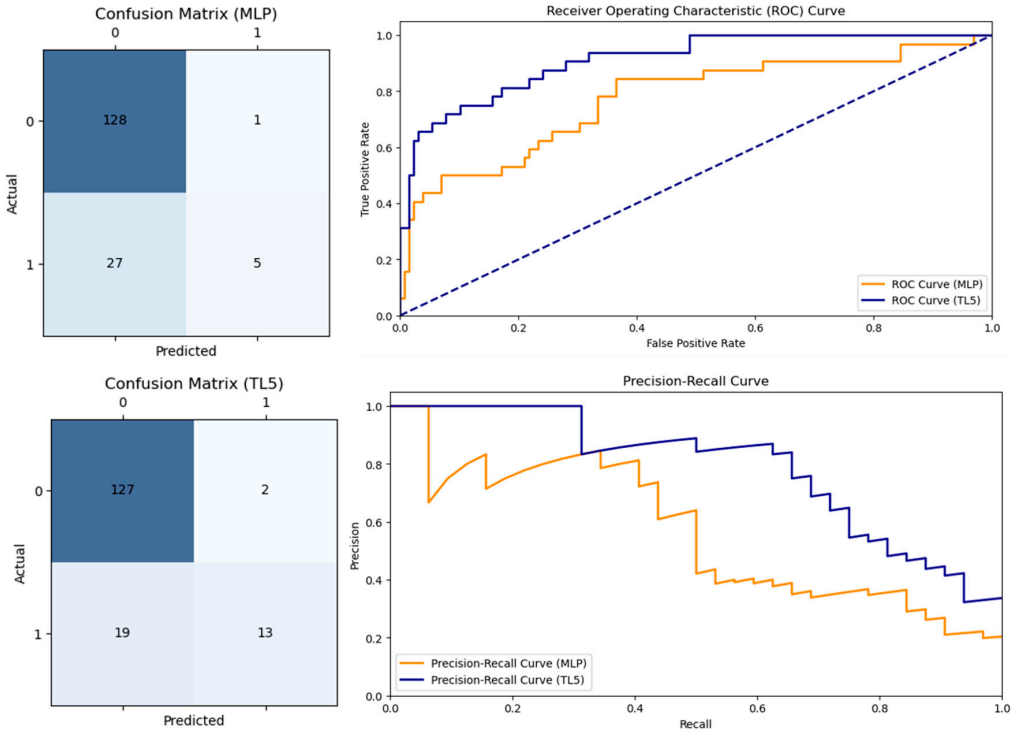

Figure A1. Confusion matrix, AUROC curve and AUPRC curve for Group 4.

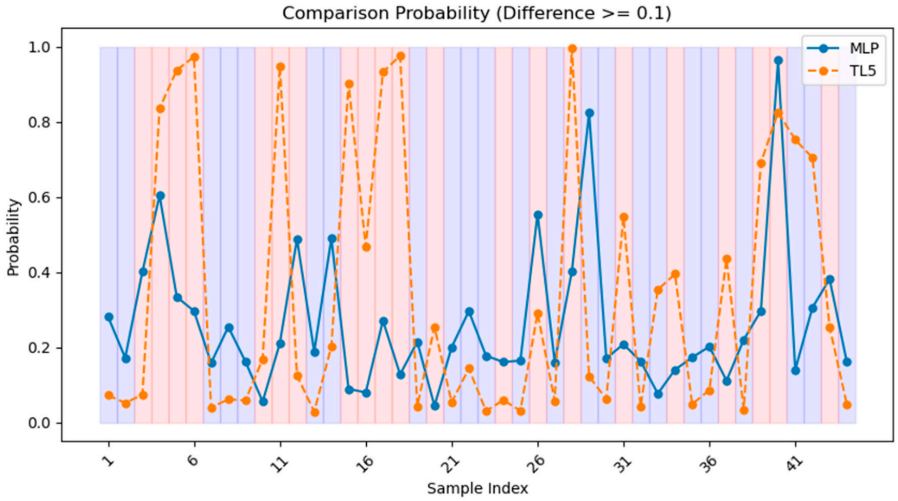

Figure A2. Comparison of probabilities between the MLP and TL5 models. Only cases with a probability difference of 0.1 or more are displayed in Group 4. Red area represents patients with ARF, while the blue area represents patients without ARF.

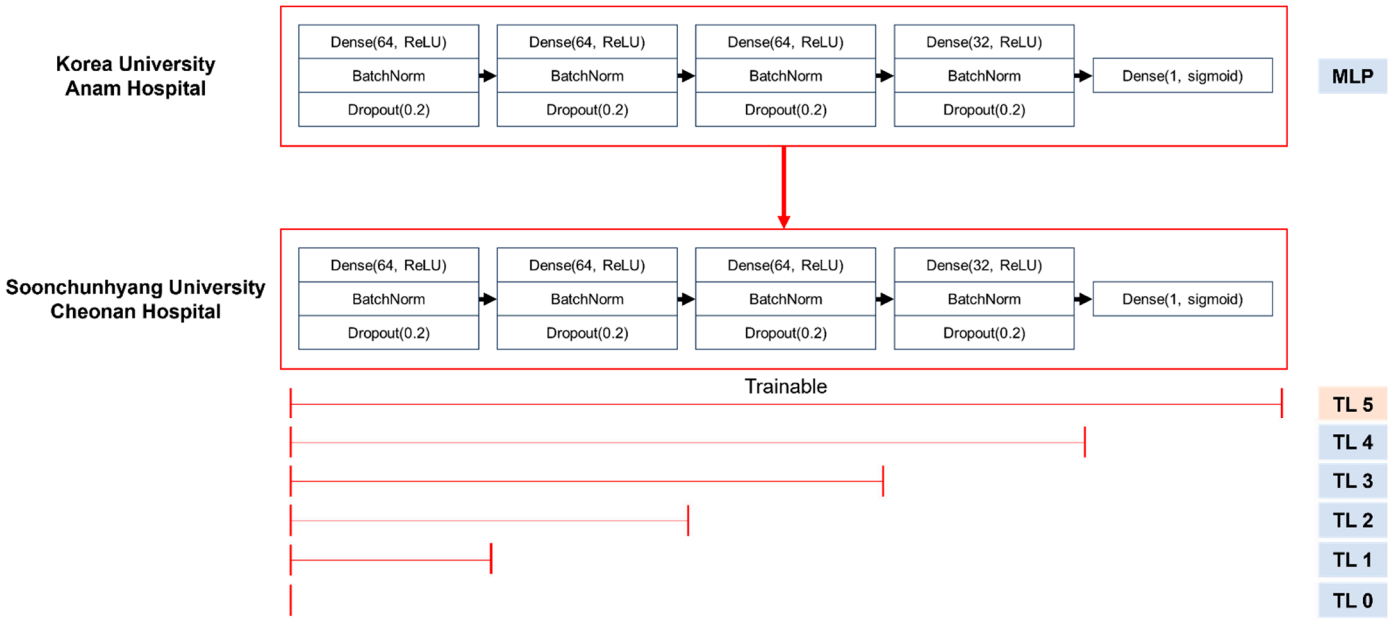

Figure A3. The structure of the MLP and TL models.

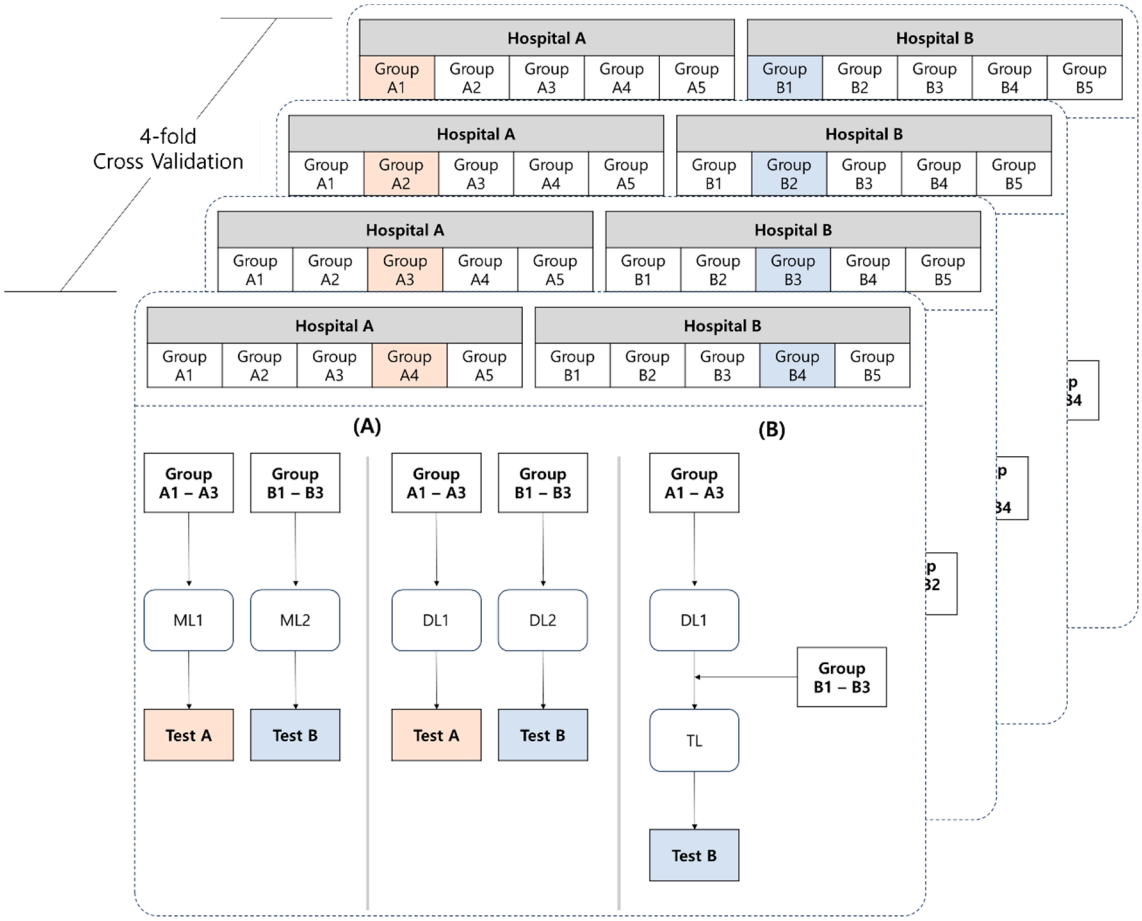

Figure A4. Model training process. (**A**) Korea University Anam Hospital, (**B**) Soonchunhyang University Cheonan Hospital.

Table A1. Abbreviation descriptions.

Category	Abbreviation	Full Form
Model	LGBM	Light Gradient-Boosting Machine
	LR	Logistic Regression
	MLP	Multi-Layer Perceptron
	RF	Random Forest
	XGB	Extreme Gradient Boosting
Metrics	AUROC	Area Under the Receiver Operating Characteristic
	AUPRC	Area Under the Precision–Recall Curve
	MCC	Matthews Correlation Coefficient
	SHAP	ShaHley Additive exPlanations
Features	BE	Base Excess
	DBP	Diastolic Blood Pressure
	GCS	Glasgow Coma Scale
	SBP	Systolic Blood Pressure
	Total CO_2	Total Carbon Dioxide
ETC	ARF	Acute Respiratory Failure
	DL	Deep Learning
	EHR	Electronic Health Records
	IRB	Institutional Review Boards
	MICE	Multiple Imputation by Chained Equations
	MV	Mechanical Ventilation
	OOD	Out-of-Distribution
	TL	Transfer Learning

Table A2. Number and proportion of missing values by feature.

Feature	Korea University Anam Hospital (*n* = 12,059)		Soonchunhyang University Cheonan Hospital (*n* = 803)	
	N	%	N	%
Age	0	0.00	0	0.00
Sex [1]	0	0.00	0	0.00
Systolic BP [2]	53,138	19.61	22	2.74
Diastolic BP	52,928	19.53	29	3.61
Respiratory	14,632	5.40	27	3.36
Heart rate	13,648	5.04	38	4.73
Serum Cr [3]	14,047	5.18	28	3.49
Hemoglobin	13,172	4.86	41	5.11
Total CO_2	11,695	4.32	65	8.09
Arterial pH	13,024	4.81	43	5.35
pCO_2	13,307	4.91	39	4.86
pO_2	13,531	4.99	43	5.35
BE [4]	13,421	4.95	43	5.35
Lactate	10,295	3.80	50	6.23
GCS [5]	251,420	92.78	24	2.99

[1] Sex, male; [2] BP, blood pressure; [3] Cr, creatinine; [4] BE, base excess; [5] GCS; Glasgow Coma Scale.

Table A3. Characteristics of pesticide exposure at Soonchunhyang University Cheonan Hospital.

Feature	N	%
Pesticide category		
Not otherwise specified	227	28.27
Glyphosate	213	26.53
Glufosinate	186	23.16
Organophosphate	90	11.21
Pyrethroid	78	9.71
Carbamate	9	1.12
Amount of ingestion		
≤50 cc	168	20.92
>50 cc, ≤100 cc	160	19.93
>100 cc, ≤200 cc	157	19.55
>200 cc, ≤300 cc	131	16.31
>300 cc	97	12.08
Unknown	90	11.21

Table A4. Model performance with Glasgow Coma Scale.

Model	AUROC		F1	
	Mean	95% CI	Mean	95% CI
LR	0.9076	0.8872–0.9279	0.6800	0.5946–0.7655
RF	0.9115	0.8687–0.9544	0.5783	0.5106–0.6460
XGB	0.9039	0.8695–0.9383	0.6316	0.5459–0.7174
LGBM	0.9056	0.8966–0.9146	0.6339	0.5401–0.7277
MLP	0.8842	0.8460–0.9225	0.6321	0.4985–0.7658

Appendix B

This appendix presents the results of additional experiments conducted to indirectly validate the utility of out-of-distribution (OOD) data in transfer learning (TL) using diverse datasets from the UCI Machine Learning Repository (https://archive.ics.uci.edu/datasets, accessed on 1 January 2024). Acknowledging potential limitations in the experimental design, we emphasize that the identification of appropriate OOD datasets is crucial. The outcomes of TL are explored across various datasets, recognizing that performance improvements may vary based on context. The utilized datasets are detailed below, and further information can be found in the UCI Machine Learning Repository:

1. Pima Indian: Smith, J.W., Everhart, J.E., Dickson, W.C., Knowler, W.C., & Johannes, R.S. (1988). https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database (accessed on 1 January 2024).
2. Cirrhosis Patient Survival Prediction: Dickson, E., Grambsch, P., Fleming, T., Fisher, L., and Langworthy, A. (2023). Cirrhosis Patient Survival Prediction. UCI Machine Learning Repository. https://doi.org/10.24432/C5R02G.
3. NHANES: National Health and Nutrition Health Survey 2013–2014 (NHANES) (2023) Age Prediction Subset. UCI Machine Learning Repository. https://doi.org/10.24432/C5BS66.
4. Wisconsin Breast Cancer: Wolberg, William, Mangasarian, Olvi, Street, Nick, and Street, W. (1995). Breast Cancer Wisconsin (Diagnostic). UCI Machine Learning Repository. https://doi.org/10.24432/C5DW2B.

5. Parkinsons Telemonitoring: Tsanas, Athanasios and Little, Max. (2009). Parkinsons Telemonitoring. UCI Machine Learning Repository. https://doi.org/10.24432/C5ZS3N.
6. CDC Diabetes Health Indicators: This dataset was released by the CDC. https://doi.org/10.24432/C53919.

We conducted multiple repetitions of the same experimental procedure across various datasets to validate the effectiveness of the proposed method. Initially, minimal data pre-processing, including handling missing values, was performed for each dataset. Subsequently, the datasets were divided into two groups to establish an OOD relationship between the majority and minority classes. One group was utilized for pre-training, while the other was employed to evaluate the proposed method. All models, including the pre-trained model, shared identical structures. Each model comprised two dense layers, with each layer incorporating batch normalization and a dropout layer with a dropout rate of 0.3. The output layer was adjusted with an appropriate activation function for regression and classification tasks. Considering data imbalance, weights were assigned to the minority class during training. For evaluation metrics, the area under the receiver operating characteristic curve (AUROC) and area under the precision–recall curve (AUPRC) were used for classification tasks, while the mean squared error (MSE) and R-squared (R^2) were employed for regression. Additionally, to account for potential variations in performance based on the method of splitting minority class data into training and test sets, we varied the random seed and repeated the process of dividing the training and test sets 300 times (7:3). The results were then averaged, and 95% confidence intervals were examined.

In the Pima Indian dataset, we stratified participants into two groups based on body mass index (BMI). The overweight group, defined as BMI 25 or above, comprised 259 individuals (40%) with diabetes, while the group with BMI below 25 included 9 individuals (7%) with diabetes (Table A5). We hypothesized a scenario where diabetes identification is targeted in the non-overweight group. We leveraged the overweight group as a pre-training model and conducted transfer learning. A comparison of the means is presented in Figure A5, and the mean values along with 95% confidence intervals for all datasets are provided in Table A12.

Table A5. Statistics of the Pima Indian dataset. "*" means statistically significant under a significance level of 0.05.

Feature	BMI $\geq$ 25 (N = 651)	BMI < 25 (N = 117)	*p*-Value
Glucose	123.36 $\pm$ 32.29	107.20 $\pm$ 26.31	<0.0001 *
Blood pressure	70.49 $\pm$ 18.02	61.39 $\pm$ 24.21	0.0002 *
Skin thickness	22.48 $\pm$ 16.06	9.71 $\pm$ 9.90	<0.0001 *
Insulin	86.90 $\pm$ 120.25	40.28 $\pm$ 70.27	<0.0001 *
DPF	0.48 $\pm$ 0.33	0.41 $\pm$ 0.31	0.0220 *
Age	33.56 $\pm$ 11.44	31.49 $\pm$ 13.34	0.1170
Pregnancies	3.0 (1.0, 6.0)	2.0 (1.0, 5.0)	0.0401 *

In the Cirrhosis Patient Survival Prediction dataset, we considered three scenarios. First, we observed a significantly higher proportion of females in the dataset. Therefore, we predicted the severity of cirrhosis in male patients. Among male patients, 127 individuals (35%) were labeled as 4, while among female patients, 7 individuals (39%) were labeled as 4. Second, we predicted the severity of cirrhosis in elderly patients aged 60 and above. Among patients under 60, 97 individuals (30%) were labeled as 4, while among patients aged 60 and above, 41 individuals (47%) were labeled as 4. Third, we predicted the severity of cirrhosis in patients who took D-penicillamine. Among patients who did not take D-penicillamine and were labeled as 4, 89 individuals (35%) were identified, while among those who took it, 55 individuals (35%) were labeled as 4. For detailed information, please refer to Table A6.

Comparisons of the means are illustrated in Figure A6. Mean values, accompanied by 95% confidence intervals for all datasets, are also detailed in Table A12.

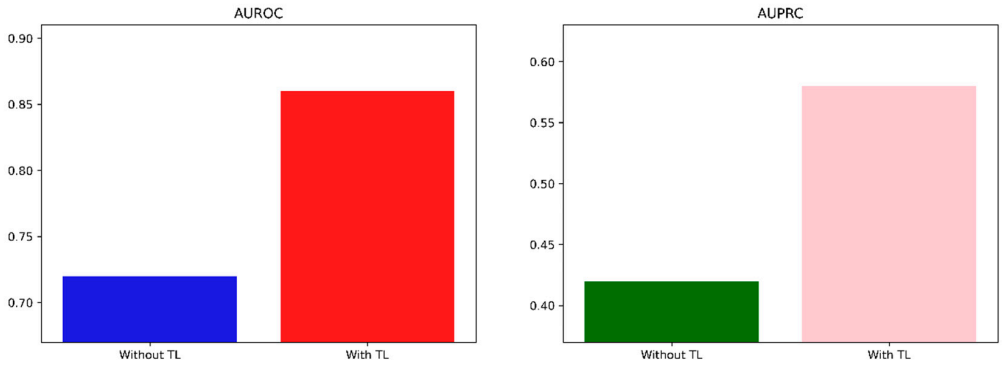

Figure A5. Comparison with and without transfer learning in the Pima Indian dataset.

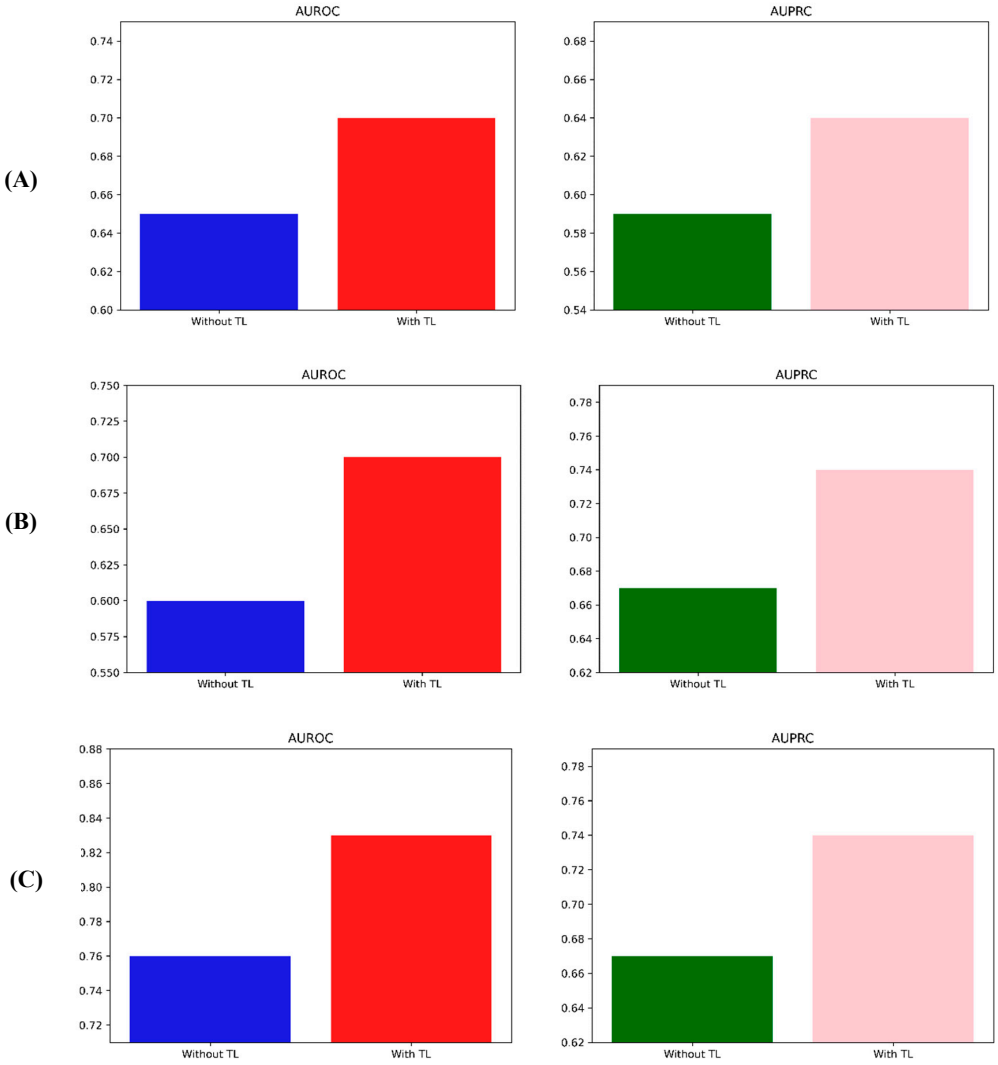

Figure A6. Comparison with and without transfer learning in the Cirrhosis Patient Survival Prediction dataset: (**A**) Male, (**B**) Elderly, (**C**) D-penicillamine.

Table A6. Statistics of the Cirrhosis Patient Survival Prediction dataset. "*" means statistically significant under a significance level of 0.05.

Feature	Dataset for Pre-Training	Dataset for Fine-Tuning	*p*-Value
	Female (N = 368)	**Male (N = 44)**	
D-penicillamine	137 (37.23)	21 (47.73)	0.2342
Ascites	21 (5.71)	3 (6.82)	1.0000
Hepatomegaly	139 (37.77)	21 (47.73)	0.2640
Spiders	86 (23.37)	4 (9.09)	0.0485 *
Edema	56 (15.22)	8 (18.18)	0.7696
Age	50.07 ± 10.25	55.75 ± 11.00	0.0020 *
Bilirubin	3.27 ± 4.62	2.87 ± 2.32	0.3426
Cholesterol	370.50 ± 238.73	362.46 ± 178.99	0.8129
Albumin	3.50 ± 0.42	3.53 ± 0.46	0.5906
Copper	90.21 ± 80.74	154.28 ± 100.67	0.0007 *
Alk_Phos	1957.83 ± 2105.05	2172.95 ± 2418.45	0.6133
SGOT	122.63 ± 57.92	121.99 ± 47.01	0.9408
Tryglicerides	123.47 ± 66.78	133.43 ± 52.17	0.3135
Platelets	259.10 ± 96.61	231.14 ± 85.23	0.0501
Prothrombin	10.71 ± 1.04	10.94 ± 0.93	0.1280
	Age < 60 (N = 324)	**Age $\geq$ 60 (N = 88)**	
D-penicillamine	120 (37.04)	38 (43.18)	0.3536
Male	27 (8.33)	17 (19.32)	0.0057 *
Ascites	13 (4.01)	11 (12.50)	0.0058 *
Hepatomegaly	126 (38.89)	34 (38.64)	1.0000
Spiders	77 (23.77)	13 (14.77)	0.0959
Edema	39 (12.04)	25 (28.41)	0.0003 *
Bilirubin	3.33 ± 4.65	2.86 ± 3.51	0.3086
Cholesterol	379.49 ± 248.56	327.96 ± 137.49	0.0392 *
Albumin	3.52 ± 0.42	3.41 ± 0.43	0.0349*
Copper	96.84 ± 86.91	101.07 ± 80.47	0.7218
Alk_Phos	2008.02 ± 2174.57	1876.12 ± 2004.30	0.6534
SGOT	124.48 ± 57.91	114.48 ± 50.97	0.1868
Tryglicerides	123.84 ± 64.71	128.27 ± 67.43	0.6603
Platelets	260.10 ± 97.20	241.24 ± 89.15	0.0916
Prothrombin	10.68 ± 0.99	10.92 ± 1.15	0.0748
	Placebo (N = 254)	**D-penicillamine (N = 158)**	
Male	23 (9.06)	21 (13.29)	0.2342
Ascites	10 (3.94)	14 (8.86)	0.0631
Hepatomegaly	87 (34.25)	73 (46.20)	0.0206 *
Spiders	45 (17.72)	45 (28.48)	0.0143 *
Edema	38 (14.96)	26 (16.46)	0.7891
Age	50.18 ± 10.10	51.47 ± 11.01	0.2324
Bilirubin	3.45 ± 4.86	2.87 ± 3.63	0.1717
Cholesterol	373.88 ± 252.48	365.01 ± 209.54	0.7474
Albumin	3.49 ± 0.41	3.52 ± 0.44	0.5485
Copper	97.65 ± 80.49	97.64 ± 90.59	0.9992
Alk_Phos	1943.01 ± 2101.69	2021.30 ± 2183.44	0.7471
SGOT	124.97 ± 58.93	120.21 ± 54.52	0.4602
Tryglicerides	125.25 ± 58.52	124.14 ± 71.54	0.8864
Platelets	254.42 ± 92.89	258.75 ± 100.32	0.6646
Prothrombin	10.78 ± 1.12	10.65 ± 0.85	0.1829

In the NHANES dataset, elderly and non-elderly individuals are labeled as 1 and 0, respectively. In this analysis, we further categorized patients into two groups: those without diabetes and those with diabetes or deemed to be in a pre-diabetic state. Among the former, 338 individuals (15%) were elderly, while among the latter, 26 individuals (33%) were elderly (Table A7). We employed the same methodology as before to predict the elderly group within the latter category. The results can be observed in Figure A7, and the 95% confidence intervals are detailed in Table A12.

Table A7. Statistics of the NHANES dataset. "*" means statistically significant under a significance level of 0.05.

Feature	No Diabetes (N = 2198)	Suspected Diabetes (N = 79)	*p*-Value
Female	1127 (51.27)	38 (48.10)	0.6631
Regular moderate-to-high-intensity exercise	1808 (82.26)	60 (75.95)	0.1986
BMI	27.83 ± 7.17	31.43 ± 8.56	0.0004 *
Glucose	98.63 ± 14.25	125.34 ± 54.05	<0.0001 *
2-h OGTT glucose	112.61 ± 42.54	180.85 ± 95.45	<0.0001 *
Insulin	11.66 ± 9.46	16.57 ± 14.53	0.0039 *

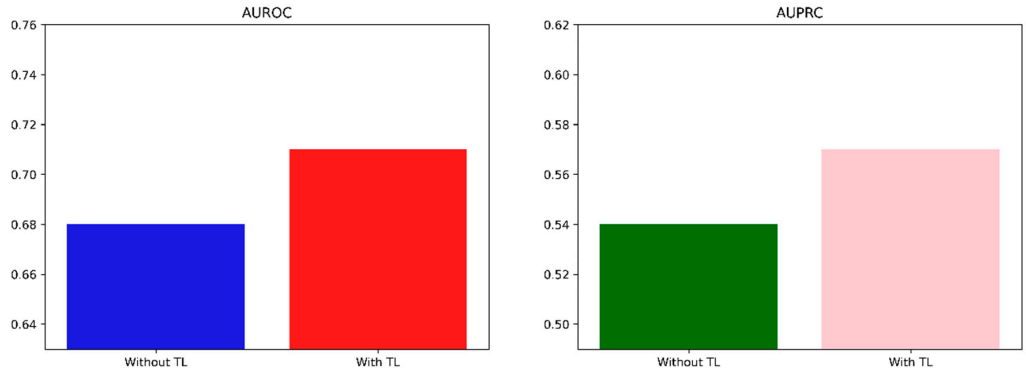

Figure A7. Comparison with and without transfer learning in the NHANES dataset.

In the Wisconsin Breast Cancer dataset, excluding the target variable, we applied the KMeans algorithm to divide the data into two groups. In Cluster 1, there were 90 individuals (20%) diagnosed with malignant tumors. In contrast, Cluster 0 contained only two individuals (0.4%), resulting in a dataset with severe class imbalance (Table A8). We undertook the task of identifying malignant tumor patients in Cluster 0. The results can be observed in Figure A8. Additionally, a comprehensive performance comparison is available in Table A12.

Table A8. Statistics of the Wisconsin Breast Cancer dataset. "*" means statistically significant under a significance level of 0.05.

Feature	Cluster 1 (N = 445)	Cluster 0 (N = 124)	*p*-Value
Radius	12.60 ± 1.92	19.62 ± 2.27	<0.0001 *
Texture	18.58 ± 4.10	21.84 ± 4.03	<0.0001 *
Perimeter	81.45 ± 13.18	129.71 ± 16.23	<0.0001 *
Area	499.67 ± 149.85	1211.94 ± 301.41	<0.0001 *
Smoothness	0.0953 ± 0.0144	0.1003 ± 0.0122	0.0001 *
Compactness	0.0928 ± 0.0460	0.1459 ± 0.0549	<0.0001 *
Concavity	0.0645 ± 0.0603	0.1760 ± 0.0802	<0.0001 *
Concave points	0.0346 ± 0.0251	0.1003 ± 0.0356	<0.0001 *
Symmetry	0.1787 ± 0.0264	0.1901 ± 0.0292	0.0001 *
Fractal dimension	0.0636 ± 0.0070	0.0599 ± 0.0064	<0.0001 *

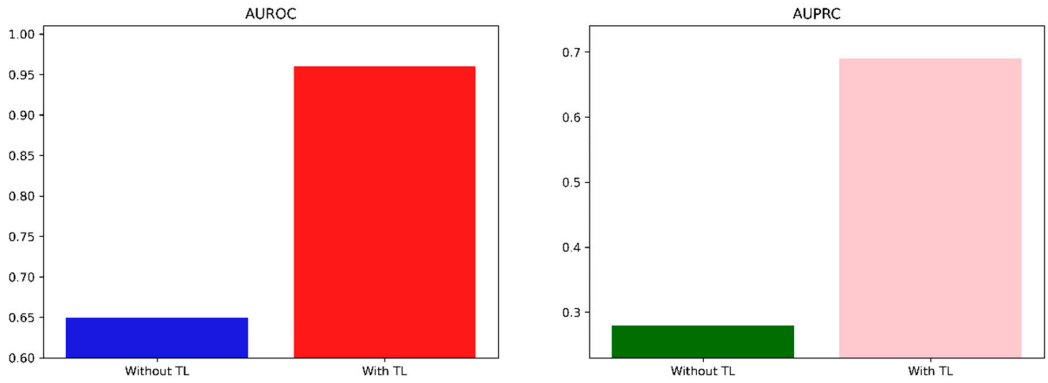

Figure A8. Comparison with and without transfer learning in the Wisconsin Breast Cancer dataset.

In the Parkinson's Telemonitoring dataset, there are two Unified Parkinson's Disease Rating Scale (UPDRS) metrics. The first is the motor UPDRS, which is also utilized as a feature, and the second is the total UPDRS. The total UPDRS is determined by considering various indicators along with the motor UPDRS. The dataset encompasses diverse data, including voice recordings, collected over six months from 44 Parkinson's patients. Drawing inspiration from the degenerative nature of Parkinson's disease, we assumed a scenario of predicting the total UPDRS in patients under the age of 60 (Table A9). To prevent the mixing of data from the same patients between the training and testing sets, we divided the data based on patients. The regression results are presented using MSE and R2. The results are depicted in Table A10.

Table A9. Statistics of the Parkinson's Telemonitoring dataset. "*" means statistically significant under a significance level of 0.05.

Feature	Age > 60 (N = 27)	Age $\leq$ 60 (N = 15)	*p*-Value
Female	8 (29.63)	6 (40.00)	0.7327
Motor UPDRS	22.93 $\pm$ 8.07	18.21 $\pm$ 7.31	<0.0001 *
Jitter (%)	0.006568 $\pm$ 0.005735	0.005370 $\pm$ 0.005323	<0.0001 *
Jitter (Abs)	0.000047 $\pm$ 0.000037	0.000039 $\pm$ 0.000033	<0.0001 *
Jitter: RAP	0.003184 $\pm$ 0.003134	0.002615 $\pm$ 0.003072	<0.0001 *
Jitter: PPQ5	0.003554 $\pm$ 0.004008	0.002752 $\pm$ 0.003078	<0.0001 *
Jitter: DDP	0.009552 $\pm$ 0.009401	0.007846 $\pm$ 0.009215	<0.0001 *
Shimmer	0.038 $\pm$ 0.028	0.027 $\pm$ 0.019	<0.0001 *
Shimmer (dB)	0.344 $\pm$ 0.252	0.249 $\pm$ 0.166	<0.0001 *
Shimmer: APQ3	0.019 $\pm$ 0.014	0.014 $\pm$ 0.010	<0.0001 *
Shimmer: APQ5	0.022 $\pm$ 0.018	0.016 $\pm$ 0.012	<0.0001 *
Shimmer: APQ11	0.030 $\pm$ 0.021	0.022 $\pm$ 0.016	<0.0001 *
Shimmered	0.057 $\pm$ 0.043	0.042 $\pm$ 0.030	<0.0001 *
NHR	0.037 $\pm$ 0.070	0.023 $\pm$ 0.030	<0.0001 *
HNR	21.21 $\pm$ 4.29	22.56 $\pm$ 4.16	<0.0001 *
RPDE	0.55 $\pm$ 0.09	0.52 $\pm$ 0.11	<0.0001 *
DFA	0.64 $\pm$ 0.07	0.67 $\pm$ 0.07	<0.0001 *
PPE	0.23 $\pm$ 0.09	0.20 $\pm$ 0.08	<0.0001 *

Table A10. Comparison with and without transfer learning in the Parkinson's Telemonitoring dataset.

Metrics	Without TL		With TL	
	Mean	95% CI	Mean	95% CI
MSE	34.08	31.54–35.61	33.40	30.55–36.25
R^2	0.36	0.28–0.45	0.42	0.35–0.49

Finally, the CDC Diabetes Health Indicators dataset includes variables with various operational definitions, and specific information can be found on the respective website. The overarching goal task is predicting diabetes status. We assumed three scenarios. The first scenario involved the prediction of diabetes in individuals who have experienced a stroke. For those without a stroke, 32,078 individuals (13%) had diabetes or pre-diabetes, while for those who had a stroke, 3268 individuals (32%) had diabetes or pre-diabetes. The second scenario involved the prediction of diabetes in individuals with coronary artery disease or heart disease. For those without the disease, 27,468 individuals (12%) had diabetes or pre-diabetes, while for those with the disease, 7878 individuals (33%) had diabetes or pre-diabetes. The third scenario involved predicting diabetes in binge drinkers. In this dataset, adult males are defined as binge drinkers if they consume 14 or more drinks per week, and adult females are defined as binge drinkers if they consume 7 or more drinks per week. For non-binge drinkers, 34,514 individuals (14%) had diabetes or pre-diabetes, while among binge drinkers, 832 individuals (6%) had diabetes or pre-diabetes. For detailed information, please refer to Table A11. Each result can be verified in Figure A9, and the comprehensive results, including 95% confidence intervals, are available in Table A12.

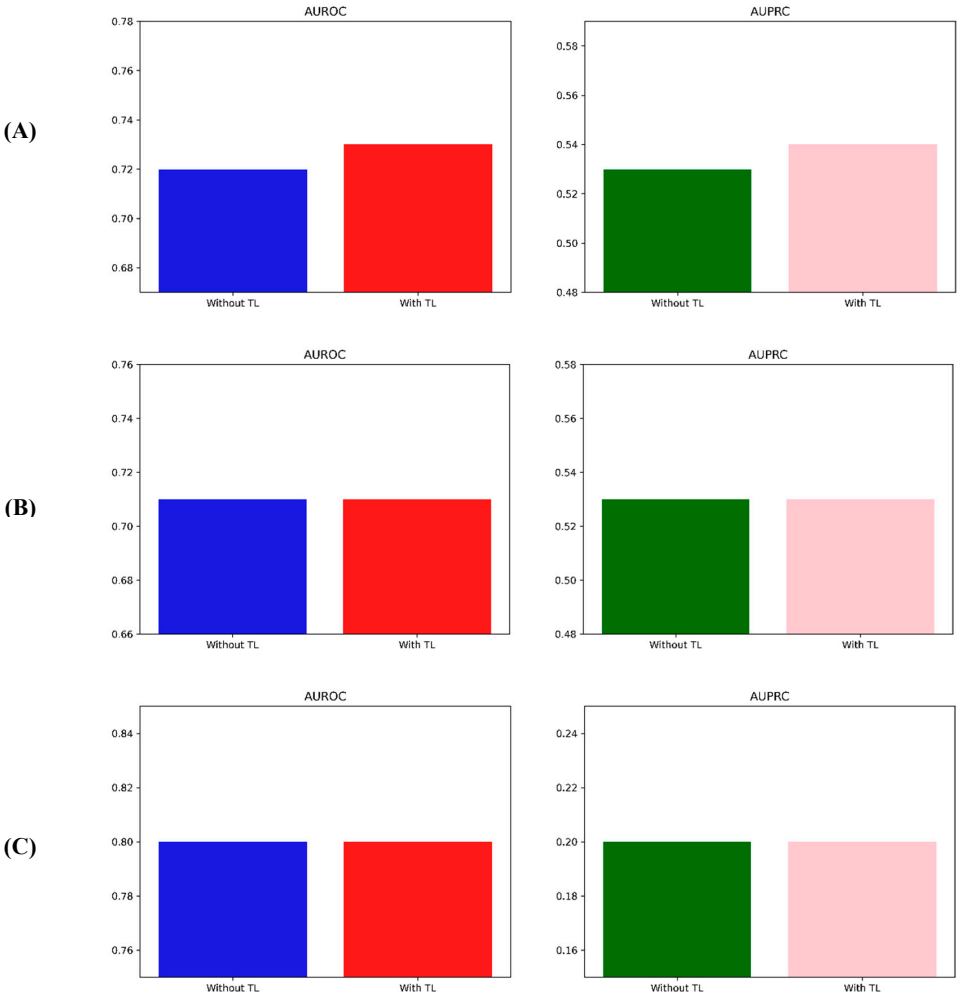

Figure A9. Comparison with and without transfer learning in the CDC Diabetes Health Indicators dataset: (**A**) Stroke, (**B**) CHD, (**C**) Binge drinker.

Table A11. Statistics of the CDC Diabetes Health Indicators dataset. "*" means statistically significant under a significance level of 0.05.

Feature	Dataset for Pre-Training	Dataset for Fine-Tuning	*p*-Value
	No Stroke (N = 243,388)	Stroke (N = 10,292)	
HighBP	101,204 (41.58)	7625 (74.09)	<0.0001 *
HighChol	100,935 (41.47)	6656 (64.67)	<0.0001 *
Smoke	106,341 (43.69)	6082 (59.09)	<0.0001 *
CHD	19,956 (8.20)	3937 (38.25)	<0.0001 *
PhysActivity	185,619 (76.26)	6301 (61.22)	<0.0001 *
Fruits	154,693 (63.56)	6205 (60.29)	<0.0001 *
Veggies	198,295 (81.47)	7546 (73.32)	<0.0001 *
Binge drinker	13,873 (5.70)	383 (3.72)	<0.0001 *
GenHlth			<0.0001 *
Excellent	44,854 (18.43)	445 (4.32)	
Very good	87,420 (35.92)	1664 (16.17)	
Good	72,473 (29.78)	3173 (30.83)	
Fair	28,591 (11.75)	2979 (28.94)	
Poor	10,050 (4.13)	2031 (19.73)	

Table A11. *Cont.*

Feature	Dataset for Pre-Training	Dataset for Fine-Tuning	*p*-Value
	No Stroke (N = 243,388)	**Stroke (N = 10,292)**	
DiffWalk	37,638 (15.46)	5037 (48.94)	<0.0001 *
Male	107,100 (44.00)	4606 (44.75)	0.1362
Age $\geq$ 60	114,696 (47.12)	7618 (74.02)	<0.0001 *
BMI	28.35 $\pm$ 6.59	29.03 $\pm$ 6.94	<0.0001 *
	No CHD (N = 229,787)	**CHD (N = 23,893)**	
HighBP	90,901 (39.56)	17,928 (75.03)	<0.0001 *
HighChol	90,838 (39.53)	16,753 (70.12)	<0.0001 *
Smoke	97,622 (42.48)	14,801 (61.95)	<0.0001 *
Stroke	6355 (2.77)	3937 (16.48)	<0.0001 *
PhysActivity	176,620 (76.86)	15,300 (64.04)	<0.0001 *
Fruits	146,450 (63.73)	14,448 (60.47)	<0.0001 *
Veggies	187,589 (81.64)	18,252 (76.39)	<0.0001 *
Binge drinker	13,408 (5.83)	848 (3.55)	<0.0001 *
GenHlth			<0.0001 *
Excellent	44,283 (19.27)	1016 (4.25)	
Very good	84,956 (36.97)	4128 (17.28)	
Good	67,732 (29.48)	7914 (33.12)	
Fair	24,842 (10.81)	6728 (28.16)	
Poor	7974 (3.47)	4107 (17.19)	
DiffWalk	32,760 (14.26)	9915 (41.5)	<0.0001 *
Male	98,018 (42.66)	13,688 (57.29)	<0.0001 *
Age $\geq$ 60	103,564 (45.07)	18,750 (78.47)	<0.0001 *
BMI	28.27 $\pm$ 6.58	29.47 $\pm$ 6.74	<0.0001 *
	Non-binge drinker (N = 239,424)	**Binge drinker (N = 14,256)**	
HighBP	102,828 (42.95)	6001 (42.09)	0.0464 *
HighChol	101,878 (42.55)	5713 (40.07)	<0.0001 *
Smoke	103,156 (43.09)	9267 (65.0)	<0.0001 *
Stroke	9909 (4.14)	383 (2.69)	<0.0001 *
CHD	23,045 (9.63)	848 (5.95)	<0.0001 *
PhysActivity	180,824 (75.52)	11,096 (77.83)	<0.0001 *
Fruits	152,849 (63.84)	8049 (56.46)	<0.0001 *
Veggies	193,792 (80.94)	12,049 (84.52)	<0.0001 *
GenHlth			<0.0001 *
Excellent	42,346 (17.69)	2953 (20.71)	
Very good	83,618 (34.92)	5466 (38.34)	
Good	71,515 (29.87)	4131 (28.98)	
Fair	30,272 (12.64)	1298 (9.10)	
Poor	11,673 (4.88)	408 (2.86)	
DiffWalk	41,100 (17.17)	1575 (11.05)	<0.0001 *
Male	105,262 (43.96)	6444 (45.20)	0.0039 *
Age $\geq$ 60	116,324 (48.58)	5990 (42.02)	<0.0001 *
BMI	28.46 $\pm$ 6.65	27.06 $\pm$ 5.79	<0.0001 *

Table A12. Comparison with and without transfer learning in all datasets.

Dataset		Metrics	Without TL		With TL	
			Mean	95% CI	Mean	95% CI
Pima Indian		AUROC	0.72	0.69–0.74	0.86	0.85–0.87
		AUPRC	0.42	0.40–0.45	0.58	0.55–0.60
Cirrhosis Patient Survival Prediction	Male	AUROC	0.65	0.63–0.66	0.70	0.69–0.72
		AUPRC	0.59	0.57–0.61	0.64	0.63–0.66
	Elderly	AUROC	0.60	0.59–0.61	0.70	0.69–0.71
		AUPRC	0.67	0.66–0.68	0.74	0.73–0.75
	D-penicillamine	AUROC	0.76	0.75–0.78	0.83	0.82–0.84
		AUPRC	0.67	0.66–0.69	0.74	0.73–0.75
NHANES		AUROC	0.68	0.67–0.69	0.71	0.70–0.72
		AUPRC	0.54	0.52–0.55	0.57	0.56–0.59
Wisconsin Breast Cancer		AUROC	0.65	0.61–0.69	0.96	0.95–0.97
		AUPRC	0.28	0.25–0.32	0.69	0.65–0.73
CDC Diabetes Health Indicators	Stroke	AUROC	0.72	0.72–0.72	0.73	0.73–0.73
		AUPRC	0.53	0.53–0.54	0.54	0.54–0.55
	CHD	AUROC	0.71	0.71–0.71	0.71	0.71–0.71
		AUPRC	0.53	0.53–0.53	0.53	0.53–0.53
	Binge drinker	AUROC	0.80	0.80–0.80	0.80	0.80–0.80
		AUPRC	0.20	0.19–0.20	0.20	0.19–0.20

When applying a consistent experimental methodology to all datasets, performance improvements were observed across almost all cases. Additionally, there was a tendency for a narrowing of the confidence interval range with the application of OOD-based TL. However, it is essential to acknowledge that we do not anticipate our proposed method to demonstrate optimal performance in every scenario, particularly in situations where either an ample amount of data is already available or where similar patterns among the data are not discernible. Nevertheless, our approach remains robust and merits consideration, especially in scenarios with limited medical data.

References

1. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 53. [CrossRef] [PubMed]
2. Lateh, M.A.; Muda, A.K.; Yusof, Z.I.M.; Muda, N.A.; Azmi, M.S. Handling a small dataset problem in prediction model by employ artificial data generation approach: A review. *J. Phys. Conf. Ser.* **2017**, *892*, 012016. [CrossRef]
3. Vapnik, V. *The Nature of Statistical Learning Theory*; Springer Science & Business Media: Berlin/Heidelberg, Germany, 1999.
4. Andonie, R. Extreme data mining: Inference from small datasets. *Int. J. Comput. Commun. Control* **2010**, *5*, 280–291. [CrossRef]
5. Tsai, T.I.; Li, D.C. Utilize bootstrap in small data set learning for pilot run modeling of manufacturing systems. *Expert Syst. Appl.* **2008**, *35*, 1293–1300. [CrossRef]
6. Niyogi, P.; Girosi, F.; Poggio, T. Incorporating prior information in machine learning by creating virtual examples. *Proc. IEEE* **1998**, *86*, 2196–2209. [CrossRef]
7. Chao, G.; Tsai, T.; Lu, T.-J.; Hsu, H.; Bao, B.; Wu, W.; Lin, M.; Lu, T. A new approach to prediction of radiotherapy of bladder cancer cells in small dataset analysis. *Expert Syst. Appl.* **2011**, *38*, 7963–7969. [CrossRef]
8. Da Silva, I.B.V.; Adeodato, P.J. PCA and Gaussian noise in MLP neural network training improve generalization in problems with small and unbalanced data sets. In Proceedings of the 2011 International Joint Conference on Neural Networks, San Jose, CA, USA, 31 July–5 August 2011; IEEE: New York, NY, USA, 2011; pp. 2664–2669.
9. Karimi, D.; Gholipour, A. Improving calibration and out-of-distribution detection in deep models for medical image segmentation. *IEEE Trans. Artif. Intell.* **2022**, *4*, 383–397. [CrossRef]

10. Major, D.; Lenis, D.; Wimmer, M.; Berg, A.; Neubauer, T.; Bühler, K. On the importance of domain awareness in classifier interpretations in medical imaging. *IEEE Trans. Med. Imag.* **2023**, *42*, 2286–2298. [CrossRef]
11. Dodge, J.; Ilharco, G.; Schwartz, R.; Farhadi, A.; Hajishirzi, H.; Smith, N. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. *arXiv* **2020**, arXiv:2002.06305.
12. Narkhede, M.V.; Bartakke, P.P.; Sutaone, M.S. A review on weight initialization strategies for neural networks. *Artif. Intell. Rev.* **2022**, *55*, 291–322. [CrossRef]
13. Izonin, I.; Roman, T. Universal intraensemble method using nonlinear AI techniques for regression modeling of small medical data sets. In *Cognitive and Soft Computing Techniques for the Analysis of Healthcare Data*; Academic Press: Cambridge, MA, USA, 2022; pp. 123–150.
14. Hekler, E.B.; Klasnja, P.; Chevance, G.; Golaszewski, N.M.; Lewis, D.; Sim, I. Why we need a small data paradigm. *BMC Med.* **2019**, *17*, 133. [CrossRef]
15. Li, D.-C.; Wu, C.-S.; Tsai, T.-I.; Chang, F.M. Using mega-fuzzification and data trend estimation in small data set learning for early FMS scheduling knowledge. *Comput. Oper. Res.* **2006**, *33*, 1857–1869. [CrossRef]
16. Shorten, C.; Khoshgoftaar, T.M. A survey on image data augmentation for deep learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
17. Mohammed, R.; Rawashdeh, J.; Abdullah, M. Machine learning with oversampling and undersampling techniques: Overview study and experimental results. In Proceedings of the 2020 11th international conference on information and communication systems (ICICS), Irbid, Jordan, 7–9 April 2020; IEEE: New York, NY, USA, 2020.
18. Zhang, Y.; Seibert, P.; Otto, A.; Raßloff, A.; Ambati, M.; Kästner, M. DA-VEGAN: Differentiably Augmenting VAE-GAN for microstructure reconstruction from extremely small data sets. *Comput. Mater. Sci.* **2024**, *232*, 112661. [CrossRef]
19. Hung, S.-K. Image Data Augmentation from Small Training Datasets Using Generative Adversarial Networks (GANs). Ph.D. Thesis, University of Essex, Colchester, UK, 2023.
20. Dou, B.; Zhu, Z.; Merkurjev, E.; Ke, L.; Chen, L.; Jiang, J.; Zhu, Y.; Liu, J.; Zhang, B.; Wei, G.-W. Machine learning methods for small data challenges in molecular science. *Chem. Rev.* **2023**, *123*, 8736–8780. [CrossRef]
21. Röglin, J.; Ziegeler, K.; Kube, J.; König, F.; Hermann, K.-G.; Ortmann, S. Improving classification results on a small medical dataset using a GAN.; An outlook for dealing with rare disease datasets. *Front. Comput. Sci.* **2022**, *4*, 858874. [CrossRef]
22. Izonin, I.; Tkachenko, R.; Bliakhar, R.; Kovac, M. An improved ANN-based sequential global-local approximation for small medical data analysis. *EAI Endorsed Trans. Pervasive Health Technol.* **2023**, *9*. [CrossRef]
23. Zhang, Y.; Zhou, D.; Hooi, B.; Wang, K. Expanding small-scale datasets with guided imagination. *arXiv* **2022**, arXiv:2211.13976.
24. Izonin, I.; Tkachenko, R.; Shakhovska, N.; Lotoshynska, N. The additive input-doubling method based on the SVR with nonlinear kernels: Small data approach. *Symmetry* **2021**, *13*, 612. [CrossRef]
25. Izonin, I.; Tkachenko, R.; Dronyuk, I.; Tkachenko, P.; Gregus, M.; Rashkevych, M. Predictive modeling based on small data in clinical medicine: RBF-based additive input-doubling method. *Math. Biosci. Eng.* **2021**, *18*, 2599–2613. [CrossRef]
26. Fanini, L.; Marchetti, G.M.; Serafeimidou, I.; Papadopoulou, O. The potential contribution of bloggers to change lifestyle and reduce plastic use and pollution: A small data approach. *Mar. Pollut. Bull.* **2021**, *169*, 112525. [CrossRef]
27. Baldominos, A.; Puello, A.; Ogul, H.; Asuroglu, T.; Colomo-Palacios, R. Predicting infections using computational intelligence–a systematic review. *IEEE Access* **2020**, *8*, 31083–31102. [CrossRef]
28. Werner, J.; Beisswanger, P.; Schürger, C.; Klaiber, M.; Theissler, A. From Data to Wisdom: A Review of Applications and Data Value in the context of Small Data. *Procedia Comput. Sci.* **2023**, *225*, 1251–1260. [CrossRef]
29. Kim, H.E.; Cosa-Linan, A.; Santhanam, N.; Jannesari, M.; Maros, M.E.; Ganslandt, T. Transfer learning for medical image classification: A literature review. *BMC Med. Imag.* **2022**, *22*, 69. [CrossRef] [PubMed]
30. Niu, S.; Liu, Y.; Wang, J.; Song, H. A decade survey of transfer learning (2010–2020). *IEEE Trans. Artif. Intell.* **2020**, *1*, 151–166. [CrossRef]
31. Kim, H.E.; Cosa-Linan, A.; Santhanam, N.; Jannesari, M.; Maros, M.E.; Ganslandt, T. Transfer learning techniques for medical image analysis: A review. *Biocybern. Biomed. Eng.* **2022**, *42*, 79–107.
32. Raghu, M.; Zhang, C.; Kleinberg, J.; Bengio, S. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*; Curran Associates: Red Hook, NY, USA, 2019.
33. Mehrtash, A.; Wells, W.M.; Tempany, C.M.; Abolmaesumi, P.; Kapur, T. Confidence calibration and predictive uncertainty estimation for deep medical image segmentation. *IEEE Trans. Med. Imag.* **2020**, *39*, 3868–3878. [CrossRef] [PubMed]
34. Lee, K.; Lee, K.; Lee, H.; Shin, J. A simple unified framework for detecting out-of-distribution samples and adversarial attacks. In *Advances in Neural Information Processing Systems 31 (NeurIPS 2018)*; Curran Associates: Red Hook, NY, USA, 2018.
35. Rajpurkar, P.; Chen, E.; Banerjee, O.; Topol, E.J. AI in health and medicine. *Nat. Med.* **2022**, *28*, 31–38. [CrossRef]
36. Cao, T.; Huang, C.-W.; Hui, D.Y.-T.; Cohen, J.P. A benchmark of medical out of distribution detection. *arXiv* **2020**, arXiv:2007.04250.
37. Cho, N.-J.; Park, S.; Lyu, J.; Lee, H.; Hong, M.; Lee, E.-Y.; Gil, H.-W. Prediction Model of Acute Respiratory Failure in Patients with Acute Pesticide Poisoning by Intentional Ingestion: Prediction of Respiratory Failure in Pesticide Intoxication (PREP) Scores in Cohort Study. *J. Clin. Med.* **2022**, *11*, 1048. [CrossRef]
38. Eddleston, M. Poisoning by pesticides. *Medicine* **2020**, *48*, 214–217. [CrossRef]
39. Eddleston, M.; Mohamed, F.; Davies, J.; Eyer, P.; Worek, F.; Sheriff, M.; Buckley, N. Respiratory failure in acute organophosphorus pesticide self-poisoning. *J. Assoc. Physicians* **2006**, *99*, 513–522. [CrossRef] [PubMed]

40. Lee, H.; Choa, M.; Han, E.; Ko, D.R.; Ko, J.; Kong, T.; Cho, J.; Chung, S.P. Causative Substance and Time of Mortality Presented to Emergency Department Following Acute Poisoning: 2014-2018 National Emergency Department Information System (NEDIS). *J. Korean Soc. Clin. Toxicol.* **2021**, *19*, 65–71. [CrossRef]
41. Kim, Y.; Chae, M.; Cho, N.; Gil, H.; Lee, H. Machine Learning-Based Prediction Models of Acute Respiratory Failure in Patients with Acute Pesticide Poisoning. *Mathematics* **2022**, *10*, 4633. [CrossRef]
42. Mera-Gaona, M.; Neumann, U.; Vargas-Canas, R.; López, D.M. Evaluating the impact of multivariate imputation by MICE in feature selection. *PLoS ONE* **2021**, *16*, e0254720. [CrossRef]
43. Yang, C.; Kors, J.A.; Ioannou, S.; John, L.H.; Markus, A.F.; Rekkas, A.; de Ridder, M.A.J.; Seinen, T.M.; Williams, R.D.; Rijnbeek, P.R. Trends in the conduct and reporting of clinical prediction model development and validation: A systematic review. *J. Am. Med. Inform. Assoc.* **2022**, *29*, 983–989. [CrossRef]
44. An, Q.; Rahman, S.; Zhou, J.; Kang, J.J. A Comprehensive Review on Machine Learning in Healthcare Industry: Classification, Restrictions, Opportunities and Challenges. *Sensors* **2023**, *23*, 4178. [CrossRef]
45. Lam, C.; Tso, C.F.; Green-Saxena, A.; Pellegrini, E.; Iqbal, Z.; Evans, D.; Hoffman, J.; Calvert, J.; Mao, Q.; Das, R. Semisupervised deep learning techniques for predicting acute respiratory distress syndrome from time-series clinical data: Model development and validation study. *JMIR Form. Res.* **2021**, *5*, e28028. [CrossRef]