A novel technique for data quality evaluation in human health PHM

John Taco[1], Brandon Foreman[2], Pradeep Kundu[1], and Jay Lee[1]

[1]*Center for Intelligent Maintenance Systems (IMS), Department of Mechanical and Materials Engineering, Cincinnati, OH, United States*

[2]*Department of Neurology and Rehabilitation Medicine, University of Cincinnati College of Medicine, University of Cincinnati Gardner Neuroscience Institute, Cincinnati, OH, United States*

tacolojo@mail.uc.edu
foremabo@ucmail.uc.edu
kundupp@ucmail.uc.edu
lj2@ucmail.uc.edu

ABSTRACT

Human healthcare data includes different signals related to the functioning of the human body such as blood pressure, intracranial pressure, heart rate, and so on. This kind of data plays an important role in ill patients in the intensive care unit. Unfortunately, the recorded data may include connection & human errors, measurement errors due to the movement of patients, and others. These errors are better known as artifacts and they should be recognized in case the data needs to be used for clinical purposes. Different methods have been proposed for artifact detection; however, the existing methods only analyze one signal at a time or rely on feature engineering. In this study, we present an alternative solution for artifact detection that analyzes multiple signals at once and does not require feature engineering. The method integrates signals such as Intracranial Pressure (ICP), Electrocardiogram (ECG), and Arterial Blood Pressure (ABP). Time raw domain data, frequency domain data, and the combination of both were studied as the input of neural network models. Four deep learning algorithms were employed: convolutional neural network (CNN), long short-term memory (LSTM), bidirectional LSTM (BiLSTM), and transformer neural network. By performing a cross-validation ensemble using a dataset of 39 patients with noisy signals, the time domain data with the LSTM model is found best in terms of accuracy with a performance of 94.81%. On the other hand, the frequency domain data with the CNN model is found best in terms of computational time. The CNN model takes 7 minutes and 3 seconds for training.

This study shows that deep learning, raw data in the frequency or time domain, and cross-validation ensemble combined have great potential for data quality evaluation in healthcare.

1. INTRODUCTION

Data quality evaluation plays a critical role in artificial intelligence algorithms' performance since they may be trained with defective information. To solve data quality problems, different preprocessing techniques have been developed as part of the general framework of machine learning.

In healthcare applications, the data can be seriously affected by different factors such as sensor errors, patient movements, measurement errors, etc. These abnormalities can cause problems such as false diagnoses and inaccurate clinical alarms. In the medical area, these errors are commonly known as artifacts and their detection is critical due to the complexity of physiological data.

Most importantly, healthcare data from the intensive care unit (ICU) has a high impact on human lives. The current death ratio in the ICU is 7% (Siddiqui, 2015). Patients with traumatic brain injuries are part of the ICU where a faster diagnosis through data-driven methods can save lives. Therefore, having good management of the data can help to make faster decisions and improve the safety and quality of patient care (Cascini et al.,2021).

Some of the most important signals in patients with traumatic brain injuries are intracranial pressure, arterial blood pressure, and electrocardiogram. Intracranial pressure (ICP) measures the pressure inside the brain. To get this measurement, it is required to create a craniotomy to insert the ICP monitor. Arterial blood pressure (ABP) measures the pressure within the arteries using an invasive catheter placed at the radial or femoral artery. An electrocardiogram (ECG) measures the activity of the heart and different values or trends in this measurement can reflect arrhythmias and other conditions. Overall, these three signals play an important role in the intensive care unit.

From the Prognostics and Health Management (PHM) perspective, the gathered data can be used for diagnosis and forecasting in different applications. In the specific case of patients with traumatic brain injuries, PHM can be used for clinical diagnosis, making it faster and more efficient. However, as mentioned before, data quality in the healthcare area is highly compromised due to people's involuntary movements or bedside care events. Hence, it is needed to evaluate the data quality first and then apply data-driven diagnosis using Machine Learning algorithms.

Different techniques can be used for this purpose such as anomaly detection, regression imputation for missing values, duplicate data identification, and so on. However, the simplicity and deployability of the methods are important factors to take into account. From the medical point of view, artifact detection significantly improves data quality. Furthermore, different kinds of artifacts are in the data and their identification is complex due to the similarity between abnormalities caused by illnesses and measurement errors. Deep learning algorithms can identify patterns in complicated data without the need for domain knowledge. They are currently used in applications such as stock market forecasting, sentiment analysis, computer vision, etc. Given the flexibility of neural networks to work in different applications, deep learning is a feasible candidate for artifact detection.

Considering that data quality evaluation plays an important role in data-driven methods due to its impact on algorithm performance, in this paper, we propose a novel method to detect artifacts by using multivariate signals and deep learning. Due to the complexity of medical data, domain knowledge is important when doing traditional artifact detection methods. Thus, the proposed deep learning methodology solves the problem by using the raw data as the input and extracting high-level features inside the neural network architecture. Neural network architectures including CNN, LSTM, BiLSTM, and Transformer neural networks are evaluated. Furthermore, a cross-validation ensemble method is performed to reduce overfitting and increase prediction power.

2. LITERATURE REVIEW.

In the last decade, data-based algorithms usage has been gaining popularity due to its flexibility to work in different industries. Deep learning, classical machine learning algorithms, and signal processing techniques can be used for data quality enhancement and predictions, which are co-dependent due to the important role of data quality on algorithm predictions. In the literature, different algorithms and methodologies for data quality evaluation using artifact and anomaly detection are available (Megjhani et al.,2019; Feng, Loy, Zhang, and Guan, 2011; Edinburgh, Smielewski, Czosnyka, Eglen, and Ercole, 2019; Subramanian et al., 2021; Gupta et al., 2022). Also, from the prediction application standpoint, different neural network applications can be found due to their capacity to recognize patterns and automatically extract features (Roy et al. 2020; Bar et al. 2015; Hussein, et al. 2019).

In this section, these two main topics are listed and described in detail: data quality evaluation methods and deep learning prediction techniques.

As for data quality techniques, Ning et al. (2017) proposed a tabular data quality improvement algorithm based on missing data imputation. The algorithm is based on a denoising auto-encoder with a high rate of missing data. Further research needs to be done for the low rate of missing data. Liu et al. (2017) proposed a data quality framework for power grid data. The proposed framework is applied to the Chinese electric power dataset due to its complexity and lack of applied data quality methods. The framework gathered real-time and historical data and gave support for data quality, storage, and management. Sulistyo et al. (2020) suggested a data enhancement method that deals with missing data. The study used Pentaho Data Integration (PDI) and the developed method was applied to a brand registration number permits dataset for Indonesian government institutions. Gu et al. (2021) developed a method for data quality improvement using crowd-based imputation and Expectation-Maximization (EM) algorithm. The model was developed in two steps, first using EM and then using crowd imputation. The results showed that the quality was improved by 10% using EM and 5% using crowd imputation. Chen et al. (2021) proposed a system and data quality enhancement methodology based on transfer learning. For validation of the proposed framework, medical data such as TwADR-L and AskAPatient have been tested and show an increase of accuracy of 6% and 12 % in each dataset respectively. Furthermore, the analysis of the two datasets showed a high correlation between data quality and machine learning algorithm performance.

As an important part of data quality evaluation, artifact detection is one of the main tasks. Then, artifact detection methods are also proposed by different authors. For instance, Megjhani et al. (2019) proposed an active learning approach for artifact detection in ICP signals. The method improved artifact labeling and showed better results than template machine, ICP stability, and threshold-based methods. Also, Feng, Loy, Zhang, and Guan (2011) developed an artifact detection solution for ICP based on signal decomposition. The method used Empirical mode decomposition (EMD) and a filtering method to detect artifacts. The filter included robust statistics as a threshold to determine the presence of an abnormality in the signal. More recently, Edinburgh, Smielewski, Czosnyka, Eglen, and Ercole (2019) proposed an artifact detection algorithm for Arterial Blood Pressure (ABP) signals. The proposed algorithm used a convolutional neural network and an auto-encoder to detect artifacts. The algorithms showed a sensitivity and specificity of around 90%. Dai et al. (2020) presented a framework that included different processing methods, machine learning techniques, and statistical techniques to specifically analyze ICP. This study also summarized previous approaches for ICP monitoring using data science and machine learning techniques, in which artifact detection plays a critical role in ICP analysis. Subramanian et al., 2021 proposed unsupervised techniques to detect artifacts in physiological data. The results show that KNN and 1 class SVM are the best options with the shortest contiguous length of artifact of 14.5039 seconds for both algorithms. Gupta et al. (2022) presented an artifact detector method based on the Savitzky-Golay filter (S-GF) with wavelet-based noise remover (WBNR). The artifact detector significantly increases the accuracy of automated detection of Bundle Branch Block (BBB), a cardiovascular complication, to 98.80%.

As for deep learning prediction techniques, Roy et al. (2020) proposed a deep learning-based model for COVID-19 diagnosis using lung ultrasonography images. The algorithm predicts the severity of the disease and the location of artifacts in the lung. Bar et al. (2015) proposed a Convolutional Neural Network (CNN) model to detect chest pathology based on 433 chest radiograph images. Krishnadas et al. (2021) suggested a malaria detection algorithm based on deep Learning and microscopic images of Peripheral Blood Smear (PCB). Both ResNet50 and DenseNet121 models were implemented and obtained an accuracy of 91.72% and 94.43% respectively. The models help with malaria detection in rural areas where the lack of health professionals increases the chance of malaria infection. Moreover, Patel, Das, Pant, and Jayasurya (2021) proposed a deep learning ensemble method to detect tuberculosis using radiographs. Weighted Ensemble and Stack generalization were used performing an accuracy of around 95.19%. Furthermore, Hussein, et al. (2019) proposed supervised and unsupervised learning approaches to perform risk stratification of tumors. For the supervised approach, the 3D convolutional neural network was used in combination with transfer learning. For improvement, the algorithm was incorporated with a computer-aided diagnosis (CAD) tool. For unsupervised learning, support vector machines were used to overcome the lack of data labeling in tumor identification. The methods were validated using pancreas and lung images. For supervised learning, a 5% improvement in accuracy with respect to deep learning methods was observed. Also, for unsupervised learning, the improvement of accuracy was found 24% with respect to SVM.

To sum up, different approaches haven been taken to perform data quality, artifact detection, and deep learning prediction techniques. This happens because data-driven strategies have been taking an essential role in the medical field and the practical usage of data improves the efficiency of healthcare services.

3. PROPOSED METHODOLOGY

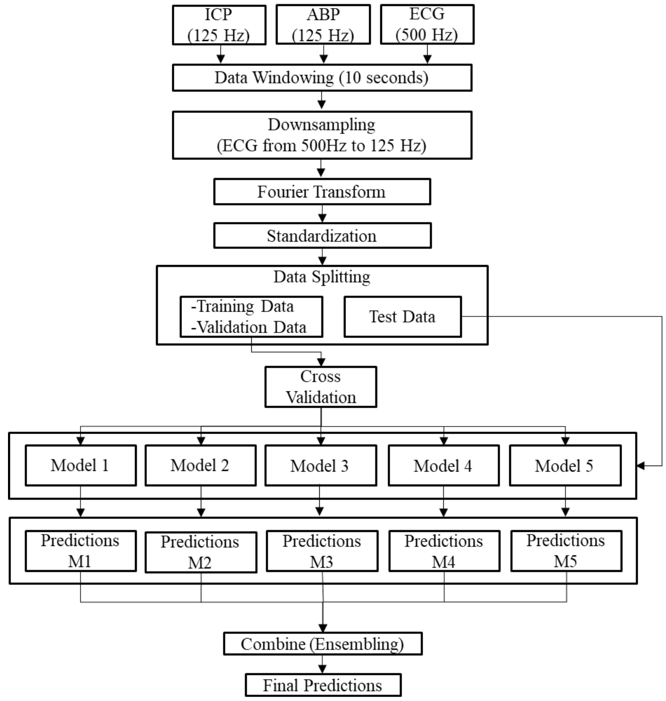

Figure 1: Data workflow

The workflow of the proposed methodology is shown in Figure 1. Further details of each step are described as follows:

3.1 Data windowing

A data windowing of 10 seconds is considered for the analysis as suggested in (Edinburgh, Smielewski, Czosnyka, Eglen, and Ercole, 2019). Given that ICP, ABP, and ECG are signals with high sampling frequency, a ten-second time window will allow the algorithms to extract all the needed information for artifact detection. For example, an ECG signal can be seen clearly within a ten-second window in Figure 2.

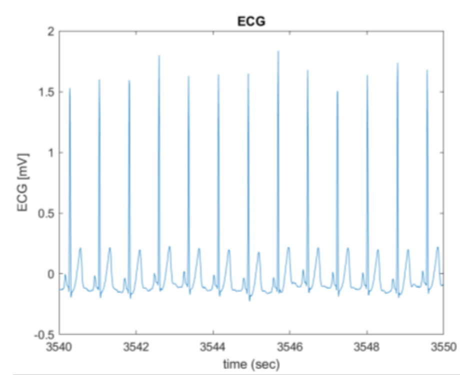

Figure 2: ECG signal

3.2 Downsampling

In medical data, different sampling frequencies appear due to the different protocols utilized by different data sources. Downsampling is needed to facilitate multivariate analysis. For multivariate time series analysis with deep learning algorithms, a matrix works as the input of the neural network architecture. Therefore, in case we have different signals as columns, all columns should have the same number of points, which means the same sampling frequency. The downsampling does not affect spikes in the ECG signal because the peaks have a frequency of about 1.5 Hz and the downsampled sampling frequency is 125 Hz. Several signals have been used to verify the effectiveness of this technique and avoid the loss of information. Figure 3 shows an illustration of the used downsampling technique.

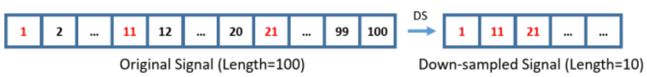

Figure 3: Downsampling

At the level matrix, we can see the effect of the downsampling in figure 4. As shown below, the signal ECG has more sample points than ICP and ABP. Then, the downsampling is applied to ECG to homogenize the three signals as represented on the right of the figure.

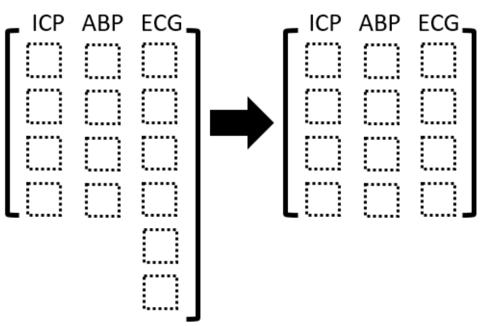

Figure 4: Data matrix after downsampling

3.3 Fourier transform

Traumatic brain injury data predominantly contains information in the frequency rather than the time domain. Since the raw data is in the time domain, frequency domain conversion is performed by feeding the Fourier transform module with time series waves of ten seconds of ICP, ABP, or ECG. By using this approach, information on the raw data such as heart rate, pulsation rate, and others are considered in the analysis. In figure 5, we can see how the time data can show different information when seen from different perspectives.

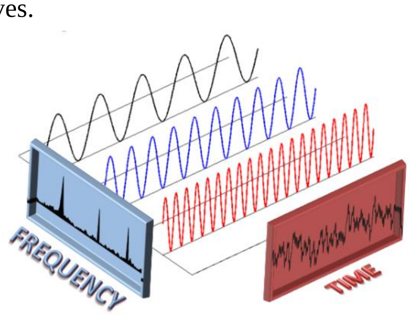

Figure 5: Fourier transform (Kalhara et al., 2017)

The electrocardiogram (ECG) has different features that can be extracted easily with a frequency-domain transformation. In figure 6, we can see an ECG sample that shows features such as RR interval, QT interval, and so on. The figure shows how the frequency between peaks, and the duration of peaks can provide useful information for the algorithm. Thus, transforming the data to the frequency domain gives us a better picture of the signal.

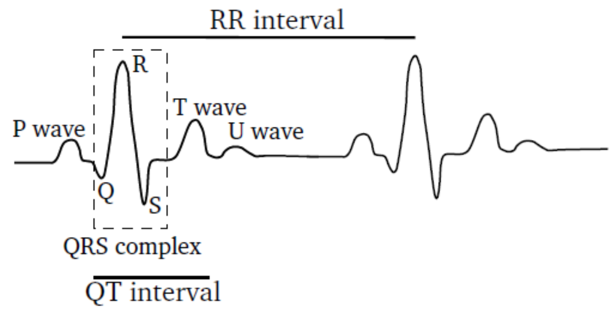

Figure 6: ECG properties (Carmona et al., 2013)

3.4 Standardization

Standardization is applied to each signal before it goes into the artificial neural network (ANN). This technique allows a faster convergence of the training process. As shown in equation 1, where σ and μ represent standard deviation (SD) and mean respectively, the Z score has been used to standardize the data. Each segment of ten seconds has been used as a reference to standardize the data.

$$\boldsymbol{Z} = \frac{\boldsymbol{x} - \boldsymbol{\sigma}}{\mu} \qquad \text{Eq. 1}$$

After applying standardization, the data will be more suitable to use since unscaled values can negatively affect the algorithm.

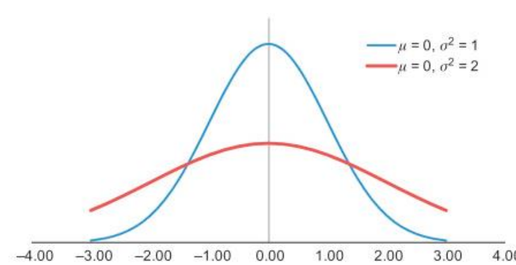

Figure 7: Standard distribution (Kissell & Poserina, 2017)

3.5 Data Splitting

To train, validate and test the algorithm, data splitting is an important step in the proposed framework. The ratio of train, validation and test data is approximately 70%-15%-15%. The selected ratio is based on the common practice of using 70% of the data for training (Kebonye, 2021) and the fact that validation and testing should have similar sizes (15%-15%). The validation data should be used for hyperparameter tuning and the test data should be used for performance evaluation. It is recommended that the test data are composed of patients that are not in the train or validation data. This way overfitting can be avoided.

Figure 8: Data splitting

3.6 Stratified Cross-validation

Stratified cross-validation allows the algorithm to have a balanced number of classes in each fold. Each fold represents a subset of the dataset. This feature helps the algorithm train with subsets of data with a similar proportion of classes. Furthermore, it enhances robustness and reduces overfitting. As shown in figure 9, the classes are well distributed in each fold, allowing the algorithm to be trained in different subsets and increasing performance due to data diversity. Each model trained in each fold can be saved and used later as a predictor. Diverse results can be. obtained by using this method, which can be combined to get a final result.

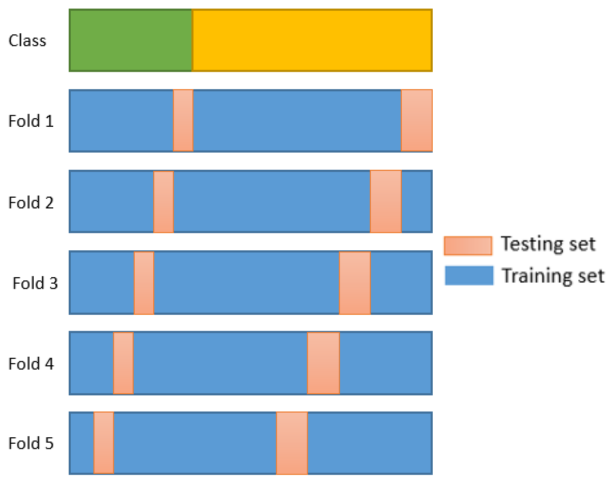

Figure 9: Stratified Cross-validation

3.7 Deep learning algorithm

Due to the complexity of medical data, deep learning algorithms are a good option to extract high-level patterns from the dataset. The final architecture for this methodology can vary depending on the performance, computational time, and hardware availability resources. Different algorithms such as recurrent neural networks, convolutional neural networks, and self-attention-based neural networks can be used to train the final model. In this study, CNN, LSTM, BiLSTM, and Transformer have been selected. CNN has been selected due to its ability to exploit spatial correlation, which can help recognize patterns between different signals when the input is a 2D matrix for multivariate time series analysis. LSTM, BiLSTM, and Transformer have been selected due to their known capacity to work with sequential data.

Also, the advantage of deep learning algorithms over traditional machine learning, in this case, is the limited use of domain knowledge. Tedious feature engineering of medical data is not required since the selected algorithms can extract features directly from the raw data. Then, the proposed method reduces the algorithm time development, and it makes the proposed framework flexible to be used in other cases. In figure 10, we can see how a multivariate time series feeds the neural network architecture. The proposed framework works as shown in figure 10, making the training and prediction more efficient and accurate.

Different sizes as the input of the neural network are presented in the presented study. The overall model architecture will not change when analyzing each case. Only the hyperparameter, which represents the size of the input signal, will change in each domain study. Also, the length of the signals either in frequency or time domain may have an impact on the model, but the main impact may be due to the valuable information in the analyzed domain.

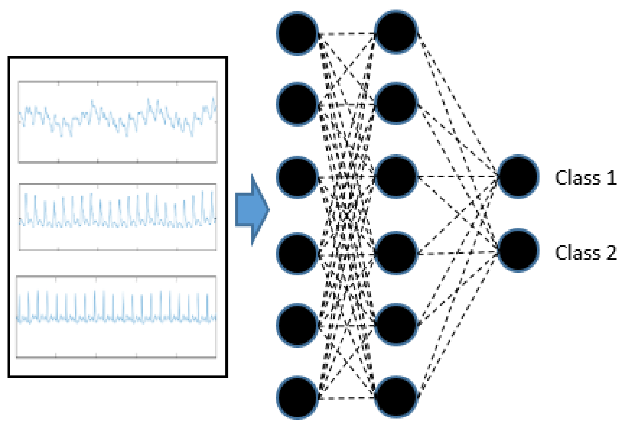

Figure 10: Artificial Neural network

- **Recurrent neural network**

A recurrent neural network is a kind of artificial neural network that allows temporal behavior connection of the data. Different algorithms are derived based on RNN, with Long short-term memory (LSTM) and Bidirectional LSTM the most popular due to their excellent results.

✓ **Long short-term memory**

Long short-term memory (LSTM) takes data as a sequence using feedback neural connections. For instance, this neural network can efficiently work with videos or audio since the sequence of the data has a meaning and the structure of the network is designed to take advantage of it. To do that, LSTM is composed of a cell, an input gate, an output gate, and a forget gate (Gers et al., 1999). Some of the successful applications of LSTM are speech recognition, sentiment analysis, video analysis, and others.

✓ **Bidirectional LSTM**

Bidirectional LSTM (Bi-LSTM) works with hidden layers from two directions (Schuster & Kuldip, 1997). This way, it can work with data from the past and the future taking advantage of valuable information from sequence data. Some of the successful applications of this neural network architecture are speech recognition, language translation, handwriting recognition, and others.

- **Convolutional neural network**

A convolutional neural network (CNN) is an artificial neural network architecture based on convolution kernels that extract high-level features from the data. It includes the following layers: input, convolution, pooling, and fully connected layer (Bezdan & Džakula, 2019). For time series, one Dimensional CNN (1D-CNN) is commonly used, and it shows good results in areas such as finances, healthcare, energy forecasting, and so on. Each layer in the architecture has a specific function as described below:

✓ Input layer: This layer contains the data in the provided form.
✓ Convolutional layer: This layer scans the input in its dimensions. Important parameters are filter and kernel size.
✓ Pooling layer: The function of this layer is to downsample the data after the convolutional operation. Typical convolutional methods are the maximum and average values.
✓ Fully connected layer: It receives a flattened input and connects it to all neurons.

- **Self-attention-based neural network**

Self-attention neural network in an artificial neural network algorithm that works based on the context of the data. One of the most popular algorithms of this type is the transformer neural network. This architecture can take advantage of information from middle points of sequence data without analyzing all the data from the beginning. Thus, it can reduce the computational training time in comparison to other recurrent neural networks (Vaswani et al.,2017). Currently, this architecture is mostly used in Natural language processing (NLP) and Computer vision (CV).

3.8 Major vote ensemble

The ensemble method helps to increase diversity within algorithms and enhance prediction power. As shown in figure 11, five models are ensembled and provide a final result. Each model represented the same architecture and algorithm, but a different subset of the dataset for training has been used. This approach is developed inside the cross-validation step, in which different subsets of the dataset are used for training. Given that different subsets are used as training data for the neural network architecture, each algorithm can give different results. The proposed methodology suggests using the average (major voting ensemble) for the predictions of all the trained algorithms, so artifact detection performance can be improved significantly.

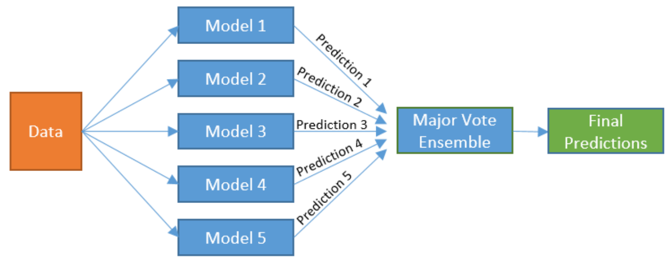

Figure 11: Major vote ensemble

4. DATA SET DESCRIPTION

The dataset was generated using 39 patients with severe traumatic brain injuries from the neuroscience intensive care unit of the University of Cincinnati Medical Center. The predominant collection duration of ICP, ABP, and ECG time series is about 5 days for each patient and a predominant subset of approximately 400 seconds has been taken for each patient for algorithm development. Each segment of data has a time window of 10 seconds and was labeled by a specialist as either artifact or non-artifact. The dataset contains three physiological signals: Intracranial pressure, Arterial Blood Pressure, and Electrocardiogram. For the purposes of this study, there were no missing values in the collected data.

Table 1: Data description

Data specifications	
Signals	ICP, ABP, ECG
Sample frequency	125 Hz (ICP, ABP), 500 Hz (ECG)
# of samples	1670
Task	Artifact Detection
Missing Values?	No
# of Patients	39
# of Artifacts	911
# Non-Artifacts	759

A segment of the three signals can be considered an artifact if one of the signals (ICP, ABP, or ECG) contains an abnormality. The sampling frequency of the signals is 125 Hz for ICP and ABP, and 500 Hz for ECG. Different sample frequencies may appear in different datasets, but the same procedure should be followed for artifact detection. The dataset contains the following columns:

- tICP: Timestamp of the ICP signal (Seconds)
- ICP: Intracranial pressure measurement (mmHg)
- tICP_start: Start measurement time of the ICP signal.

- tABP: Timestamp of the ABP signal (Seconds)
- ABP: Arterial blood pressure measurement (mmHg)
- tABP_start: Start measurement time of the ABP signal (Sec.)
- tECG: Timestamp of the ICP signal (Sec.)
- ECG: Electrocardiogram measurement (mV).
- tECG_start: Start measurement time of the ECG signal.
- manual_label: Manual label (Artifact=1, Non-Artifact=1)
- AbnormalSignal: Cause of the artifact (1: ICP, 2: ABP, 3:ECG)
- Patient: Patient number identification number (ID).

The percentage of artifacts and non-artifacts in the dataset is 55% and 45% respectively. In figure 12, we can see one normal signal (non-artifact) and two samples of artifacts due to ICP and ECG. The first plot shows a seasonal signal with a constant trend in ICP, ABP, and ECG. Given that the three signals show normal behavior; this segment is considered a non-artifact. The second plot shows disruptions in ICP, ABP and ECG. Given the abnormalities in ICP, ABP, and ECG, this signal is considered an artifact.

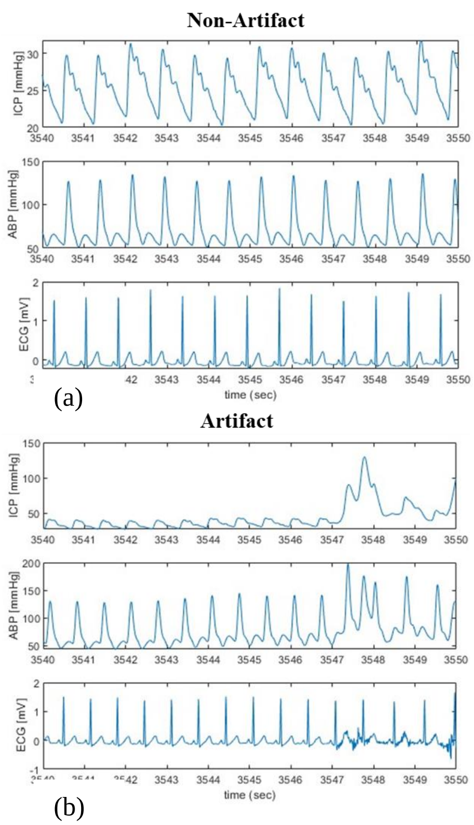

Figure 12: Signal samples (a) non-artifact (b) artifact

5. PREDICTIONS & RESULTS

The proposed methodology for artifact detection is validated on the test dataset that contains 270 samples including artifacts and non-artifacts. The results are given based on accuracy, True positive (TP), False positive (FP), False negative (FN), True negative (TN), and GPU time. The main three parameters for this study are accuracy, false negative, and GPU time. False negative is considered since a high false negative ratio implies that the algorithm allows artifact signals to pass as normal signals decreasing the overall quality of the data.

Furthermore, three domains have been considered as the input of the neural network as follows: time domain, Fourier transformation, and time and Fourier transform combined. All domains are tested in the same neural network architecture of each deep learning algorithm (CNN, LSTM, BiLSTM, or Transformer)

- **Neural network architectures**

The neural network architectures have been defined by using the Optuna optimizer. The final architectures for each algorithm are defined in table 2 and 3:

Table 2: CNN, LSTM & BiLSTM architectures

	CNN	LSTM	BiLSTM
# Input nodes	256	50	96
# Hidden layers	3	1	2
# nodes in hidden layers (HL)	HL1: 64, HL2: 32 HL3: 32	HL1: 50	HL1: 32, HL2: 32
Learning rate	0.001	0.001	0.001
Batch Size	32	32	32
# of Epochs	200	200	200

Table 3 describes the architecture of the transformer neural network. Different hyperparameters are defined in comparison to the CNN, LSTM, and Bilstm algorithms

Table 3: Transformer architecture

	Head Size	# of heads	# of blocks	Multilayer Perceptron Units
Transformer	256	2	2	128

- **Time domain data approach**

For the time domain approach, the input for all networks is as shown in figure 13. The ECG signal has been donwsampled so it has1250 samples in 10 seconds as the other signals.

Figure 13: Time domain matrix input

As shown in table 4, LSTM has the best accuracy (94.81%) among the algorithms with time-domain data. In terms of computational time (GPU time), all the algorithms but CNN have a training time of more than one hour. The lowest False negative ratio is obtained by CNN.

Table 4: Time domain results

	CNN	LSTM	BiLSTM	Transformer
Input	Time raw data			
Accuracy	87.78%	94.81%	91.11%	80.00%
TP	91.55%	90.14%	90.85%	83.10%
FP	16.41%	0.00%	8.59%	23.44%
FN	8.45%	9.86%	9.15%	16.90%
TN	83.59%	100.00%	91.41%	23.44%
GPU time	7min 24s	1h 4min	3h 26min	1h 22min

- **Frequency domain data approach**

For the frequency domain approach, Fourier transformation has been taken for all signals, and the input of all algorithms is shown in figure 14. In this case, half of the signal length is taken because for the Fourier transform half of the signal is mirrored because of conjugate symmetry.

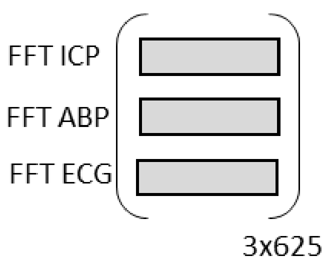

Figure 14: Frequency domain matrix input

As shown in table 5, LSTM and BiLSTM have the best accuracy (92.96%) within the algorithms with frequency domain data. CNN has the lowest computational time (7 min 3 sec) for training. LSTM has the lowest False-negative ratio.

Table 5: Frequency domain results

	CNN	LSTM	BiLSTM	Transformer
Input	Fourier transformation data (Frequency domain)			
Accuracy	92.22%	92.96%	92.96%	91.85%
TP	88.79%	96.48%	95.77%	90.14%
FP	3.91%	10.94%	10.16%	6.25%
FN	11.27%	3.52%	4.23%	9.86%
TN	96.09%	89.06%	89.84%	93.75%
GPU time	7min 3s	1h	3h 31min	1h 20min

- **Time and frequency domain data approach**

For the time-frequency domain combined, the input for all algorithms is shown in figure 15. It can be seen that FFT and the time domain data have the same length. This is because, in the frequency domain, the whole length of the signal without deleting the mirrored part was taken.

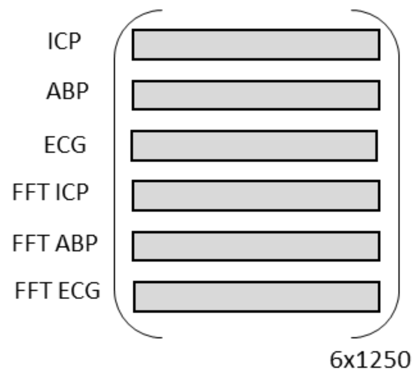

Figure 15: Time-Frequency domain matrix input

As shown in table 6, BiLSTM has the best accuracy (92.59%) among the algorithms with the time-frequency domain features. On the other hand, CNN has the lowest computational time (8 min) and the lowest false-negative ratio.

Table 6: Time-Frequency domain results

	CNN	LSTM	BiLSTM	Transformer
Input	Frequency and Time domain data			
Accuracy	91.85%	91.85%	92.59%	90.74%
TP	95.07%	92.96%	92.96%	87.32%
FP	11.72%	9.38%	7.81%	5.47%
FN	4.93%	7.04%	7.04%	12.68%
TN	88.28%	90.63%	92.19%	94.53%
GPU time	8min	57min16s	3h 30min	1h 21min

Overall, CNN has the lowest computational time and LSTM & BiLSTM have the best accuracies. We can see the difference between the performance in Figures 16 and 17 for accuracy and computational time respectively.

In figure 13, we can see that LSTM in the time domain has the best performance. For the other algorithms, the best performance occurs at the frequency domain (Fourier transform)

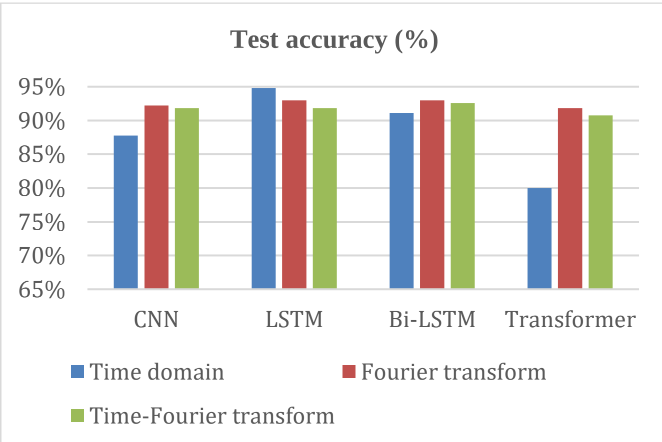

Figure 16: Accuracy of different deep learning models utilizing different types of data

In figure 14, we can see that, by far, the lowest computational time belongs to the CNN model. The training time for CNN is 7 min and for the other algorithms is between 1-4 hours. This suggests a much better use of hardware resources from the CNN model.

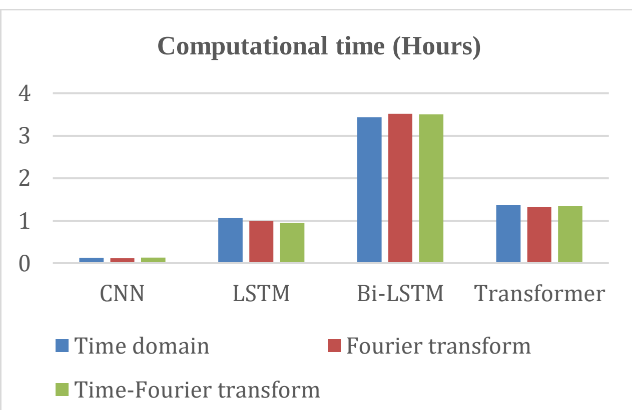

Figure 17: Computational time calculated using different deep learning approaches

6. CONCLUSIONS AND DISCUSSION

In this study, a framework for data quality improvement using artifact detection has been developed. The framework was developed to leverage multiple signals through multivariate analysis and to provide artifact label annotation using both time and frequency domain data from patients with severe traumatic brain injuries in the neuroscience intensive care unit. The proposed framework establishes a data quality improvement PHM methodology that can be used in other areas with high-frequency data.

In terms of performance, LSTM and Bi-LSTM have the best accuracy results. For LSMT, the average accuracy within all domains is 93.2% and the best accuracy is 94.81 in the time domain. For BiLSMT, the average accuracy within all domains is 92.22% and the best accuracy is 92.96% in the frequency domain.

In terms of computational time, CNN has, by far, the best result. It only takes 7-8 minutes to train the algorithm whereas for the other algorithms it takes between 1-3 hours.

In terms of the input of the neural network architecture, the frequency domain has overall the best results. Three out of four algorithms show better performance in the frequency domain than in the time and time & frequency domain approach.

In terms of false-negative rate, the results vary for each algorithm depending on the input of the neural network (Time domain, Frequency domain, or Time & Frequency domain). The best false-negative ratio is 3.52% which comes from LSTM using frequency domain input for the neural network.

In terms of preprocessing, the most feasible solutions are the frequency and time domain approaches separately. The combined time & frequency domain approach can cause some complexity in the preprocessing part of the algorithm because it includes merging the time and frequency domain data in the same neural network input.

Overall, the LSTM model with time domain data as the input has the best performance in terms of accuracy (94.81 %) and, when using time domain data, it has the best false-negative ratio (3.52%). In terms of computational time, the CNN model has the best training time (7 minutes), which means that the use of hardware resources is the best among the other algorithms. Importantly, the accuracy and false-negative rate of the CNN model using time-frequency domain information is comparable (91.85% accuracy, 4.93% false-negative rate) to the optimal model which makes this an attractive option for online processing.

In conclusion, the best options in terms of accuracy and computational time can be used under different requirement conditions. For instance, the CNN model can be used for an active learning approach in which computational time can be an important factor to consider. On the other hand, the LSTM model can be used in different conditions when retraining the model is not frequently required. In general, the two options have good accuracy. For CNN, the best accuracy is 92.22% in the frequency domain and for LSTM, the best accuracy is 94.81% in the time domain. The selection of which algorithm can be feasible depends on the user requirements.

REFERENCES

Patel, M., Das, A., Pant, V. and M, J. (2021) Detection of Tuberculosis in Radiographs using Deep Learning-based Ensemble Methods. *Smart Technologies, Communication and Robotics (STCR),* pp. 1-7, doi: 10.1109/STCR51658.2021.9588936.

Hussein, S., Kandel, P., Bolan, C. W., Wallace, M. B., and Bagci, U. (2019) Lung and Pancreatic Tumor Characterization in the Deep Learning Era: Novel Supervised and Unsupervised Learning Approaches, *IEEE Transactions on Medical Imaging*, vol. 38, no. 8, pp. 1777-1787, doi: 10.1109/TMI.2019.2894349.

Megjhani, M., Alkhachroum, A., Terilli, K., Ford, J., Rubinos, C., Kromm, J., Wallace, B. K., Connolly, E. S., Roh, D., Agarwal, S., Claassen, J., Padmanabhan, R., Hu, X., & Park, S. (2019). An active learning framework for enhancing identification of non-artifactual intracranial pressure waveforms. *Physiological measurement*, 40(1), 015002. doi: https://doi.org/10.1088/1361-6579/aaf979

Feng, M., Loy, L. Y., Zhang, F., & Guan, C. (2011). Artifact removal for intracranial pressure monitoring signals: a robust solution with signal decomposition. *Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 797–801), 2011. doi: https://doi.org/10.1109/IEMBS.2011.6090182

Edinburgh, T., Smielewski, P., Czosnyka, M., Cabeleira, M., Eglen, S. J., & Ercole, A. (2021). DeepClean: Self-Supervised Artefact Rejection for Intensive Care Waveform Data Using Deep Generative Learning. *Acta neurochirurgica. Supplement*, 131, 235–241. https://doi.org/10.1007/978-3-030-59436-7_45

Roy, S., et al. (2020). Deep Learning for Classification and Localization of COVID-19 Markers in Point-of-Care Lung Ultrasound, *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2676-2687, doi: 10.1109/TMI.2020.2994459.

Bar, Y., Diamant, I., Wolf, L., Lieberman, S., Konen, E. and Greenspan, H. Chest pathology detection using deep learning with non-medical training, *IEEE 12th International Symposium on Biomedical Imaging (ISBI)*, (pp. 294-297), 2015, doi: 10.1109/ISBI.2015.7163871.

Krishnadas, P. and Sampathila, N. Automated Detection of Malaria implemented by Deep Learning in Pytorch, *IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, (pp. 01-05), 2021, doi: 10.1109/CONECCT52877.2021.9622608.

Ning, X., Xu, Y., Gao, X. and Li, Y. Missing data of quality inspection imputation algorithm base on stacked denoising auto-encoder, *IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, pp. 84-88, 2017, doi: 10.1109/ICBDA.2017.8078781.

Liu, H., Huang, F., Li, H., Liu, W. and Wang, T. A Big Data Framework for Electric Power Data Quality Assessment, *14th Web Information Systems and Applications Conference (WISA)*, pp. 289-292, 2017, doi: 10.1109/WISA.2017.29.

Sulistyo, H. A, Kusumasari, T. F., and Alam, T. N. Implementation of Data Cleansing Null Method for Data Quality Management Dashboard using Pentaho Data Integration, *3rd International Conference on Information and Communications Technology (ICOIACT)*, pp. 12-16, 2020, doi: 10.1109/ICOIACT50329.2020.9332030.

Gu, B., Li, Z., Liu, A., Xu, J., Zhao, L., and Zhou, X. (2020). Improving the Quality of Web-Based Data Imputation With Crowd Intervention, *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 6, pp. 2534-2547, doi: 10.1109/TKDE.2019.2954087.

Chen, H., Chen, J. and Ding, J. (2021). Data Evaluation and Enhancement for Quality Improvement of Machine Learning, *IEEE Transactions on Reliability*, vol. 70, no. 2, pp. 831-847, doi: 10.1109/TR.2021.3070863.

Carmona, N., Rúa-Seoane, J. C., Elorza, J., de Pipaón, E. S., Palacios, C., Bragard, J., Aging of ECG characteristics over a five year period, *Computing in Cardiology*, pp. 1031-1034, 2013.

Siddiqui S. (2015). Mortality profile across our Intensive Care Units: A 5-year database report from a Singapore restructured hospital. *Indian journal of critical care medicine : peer-reviewed, official publication of Indian Society of Critical Care Medicine*, 19(12), 726–727. doi: https://doi.org/10.4103/0972-5229.171401

P. G. Kalhara, V. D. Jayasinghearachchi, A. H. A. T. Dias, V. C. Ratnayake, C. Jayawardena and N. Kuruwitaarachchi, TreeSpirit: Illegal logging detection and alerting system using audio identification over an IoT network, *11th International Conference on Software, Knowledge, Information Management and Applications (SKIMA)*, pp. 1-7, 2017 doi: 10.1109/SKIMA.2017.8294127.

Kissell, R. & Poserina, J. (2017). *Optimal Sports Math, Statistics, and Fantasy*. Academic Press.

Gers, F. A., Schmidhuber, J., Cummins, F. Learning to Forget: Continual Prediction with LSTM, *Neural Computation*, vol. 12, pp. 2451-2471, 1999, doi: 10.1109/ICBDA.2017.8078781.

Schuster, M., and Kuldip, P. (1997). Bidirectional recurrent neural networks, IEEE Transactions on Signal Processing, vol 45, pp 2673 – 2681, doi :10.1109/78.650093.

Bezdan, T, and Džakula, N. B. Convolutional Neural Network Layers and Architectures, *International Scientific Conference On Information Technology and Data Related Research*, pp 445-451 ,2019 , doi: 10.15308/Sinteza-2019-445-451.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., Polosukhin, I. Attention Is All You Need, *Proceedings of the 31st International Conference on Neural Information Processing System*, pp. 6000–6010, 2017 doi: https://doi.org/10.48550/arXiv.1706.03762.

Dai, H., Jia, X., Pahren, L., Lee, J., & Foreman, B. (2020). Intracranial Pressure Monitoring Signals After Traumatic Brain Injury: A Narrative Overview and Conceptual Data Science Framework. *Frontiers in neurology*, vol 11, 959. https://doi.org/10.3389/fneur.2020.00959

Cascini, F., Santaroni, F., Lanzetti, R., Failla, G., Gentili, A., & Ricciardi, W. (2021). Developing a Data-Driven Approach in Order to Improve the Safety and Quality of Patient Care. Frontiers in public health, 9, 667819. https://doi.org/10.3389/fpubh.2021.667819

Kebonye, N. M. (2021). Exploring the novel support points-based split method on a soil dataset. Measurement: Journal of the International Measurement Confederation, 186(June), 110131. https://doi.org/10.1016/j.measurement.2021.110131

Gupta, K., Bajaj, V., & Ansari, I. A. (2022). An Improved Deep Learning Model for Automated Detection of BBB Using S-T Spectrograms of Smoothed VCG Signal. IEEE Sensors Journal, 22(9), 8830–8837. https://doi.org/10.1109/JSEN.2022.3162022

Subramanian, S., Tseng, B., Barbieri, R., & Brown, E. N. (2021). Unsupervised Machine Learning Methods for Artifact Removal in Electrodermal Activity. Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS, 399–402. https://doi.org/10.1109/EMBC46164.2021.9630535