# CONTENTS:

# TABLE OF DEFINITIONS:

| **Features** | **Description** |
|---|---|
| **Study No** | Number associated with each patient within the dataset. |
| **Hospital No** | Unique identifier for each patient in the study. |
| **Sex** | Gender of the neonate, either female or male. |
| **GA (w)** | Gestational Age, measured in weeks. |
| **BW (g)** | Body Weight of the neonate, measured in grams. |
| **BW z-score** | Standardized z-score for Body Weight. |
| **Diagnosis** | Medical diagnosis of the neonate. |
| **BPD** | Bronchopulmonary Dysplasia presence, either 0 or 1. |
| **DOL** | Day Of Life, measured in days. |
| **Time** | The amount of time neonate spends on ventilator recorded hourly. |
| **Ventilation Mode** | The mode of respiratory support used. |
| **PIP** | Peak Inspiratory Pressure, a measure used in ventilator settings. |
| **PEEP** | Positive End-Expiratory Pressure, a measure used in ventilator settings. |
| **MAP** | Mean Airway Pressure, a measure used in mechanical ventilation. |
| **Vte** | Tidal Volume Exhaled, measures the amount of air the neonate exhales. |
| **Itime** | Inspiratory time, duration of the inhalation phase during mechanical ventilation. |
| **Rate** | Respiration rate, number of breaths per minute. |
| **Trigger rate** | The rate at which the ventilator is triggered by the neonate's spontaneous efforts. |
| **Compliance** | Respiratory system compliance, indicating the change in volume for each unit change in pressure. |
| **Resistance** | Airway resistance, indicating the resistance to airflow in the respiratory airways. |
| **FiO2** | Fraction of Inspired Oxygen, the concentration of oxygen in the gas mixture being inhaled by the neonate. |

# ABSTRACT:

Neonatal intensive care requires accurate and real-time monitoring of end-tidal carbon dioxide (etCO2) levels to provide optimal respiratory support for vulnerable newborns. Traditional methods for monitoring CO2 are limited by accuracy and invasiveness, emphasising the urgent need for precise, non-invasive, and interpretable solutions. In this study, machine learning (ML) and explainable artificial intelligence (XAI) were explored to improve the prediction of etCO2 levels in neonates. This approach combines the predictive power of ML with the transparency and interpretability of XAI. By carefully analysing data and engineering features and evaluating models, the XGBoost model was found to be the best-performing model with a MAPE of 4.8% and no observable overfitting. The XGBoost model outperformed traditional ML methods like linear regression by five times. These findings suggest that with further research, this model could potentially be implemented into the clinics to enhance neonatal care.

# 1 INTRODUCTION:

### 1.1 Project Motivations

Neonatal intensive care presents challenges, particularly in providing respiratory support for newborns. One of the most critical yet problematic aspects is the accurate and continuous monitoring of carbon dioxide (CO2) levels, which has a direct impact on the well-being and recovery of neonates. However, traditional methods of CO2 monitoring suffer from limitations in accuracy, invasiveness, and the lack of real-time data processing capabilities. To address these gaps in neonatal care, this study explores the application of machine learning (ML) and explainable artificial intelligence (XAI) technologies. By utilizing the predictive power and interpretability of these advanced techniques, this project aims to introduce a more effective approach to predicting etCO2 levels.

### 1.2 Aims and Objectives

The main objective of this study is to use Machine Learning (ML) and Explainable AI (XAI) to predict future levels of CO2 in neonates accurately, reliably, and efficiently. To achieve this, the study has set out specific objectives which include:

- Developing various ML models, such as tree-based and deep learning, to predict future etCO2 levels in neonates.
- Assessing the impact of data pre-processing, feature engineering, and model optimisation on these models to increase their complexity and performance.
- Evaluating the performance of various ML models.
- Identifying the most accurate and clinically applicable ML model for real-time etCO2 monitoring in neonatal care.
- Successfully implementing XAI based on the best-performing model in this study.
- Exploring the potential of integrating the best-performing ML model into clinical practice to improve neonatal care outcomes.

### 1.3 Report Structure

The report is structured to provide a comprehensive overview of the research:

**Section 1** delves into the background, highlighting the significance of respiratory support and CO2 monitoring in neonatal care and discussing the limitations of current monitoring techniques.

**Section 2** outlines the methodologies used and evaluates various ML models.

**Section 3** presents the model development process, detailing the data pre-processing, feature selection, feature engineering, and data augmentation.

**Section 4** explains the training process of all the models used in the study.

**Section 5** analyses the results, comparing the effectiveness of different models and discussing the implications of our findings for neonatal care.

**Section 6** discusses professionalism and responsibility in the context of applying ML to healthcare.

# SECTION 1

In neonatal care, respiratory support is critical for the well-being of newborns, particularly those born prematurely with respiratory conditions. Predicting carbon dioxide (CO2) levels accurately is essential for effective respiratory management in neonates. Since CO2 has a significant impact on respiratory outcomes, continuous monitoring of its levels is vital. However, existing monitoring methods have limitations, highlighting the need for innovative approaches such as machine learning (ML) models for CO2 prediction. This section aims to introduce neonatal care, explain its limitations, and explore possible improvements.

# 2 BACKGROUNDS:

### 2.1 Neonatal Respiratory Support

Neonatal respiratory support is critical to caring for newborns, particularly those born prematurely with respiratory conditions. The primary objective of respiratory support is to ensure that the baby receives enough oxygen and ventilation while minimising damage to the lungs and improving long-term outcomes. There are two main types of respiratory support used in neonatal care: noninvasive and invasive methods [1].

Non-invasive respiratory is a method that uses a mask or nasal prongs in the noted mouth or nose to provide respiratory support [1]. Non-invasive methods, such as nasal continuous positive airway pressure (nCPAP)), have become increasingly popular in neonatal care. This method provides adequate respiratory support while minimising the risks associated with invasive procedures. They are also referred to as initial interventions for preterm infants, as they help maintain lung volume, improve oxygenation, and reduce the need for intubation and mechanical ventilation. Hence decreasing the risk of lung injury [2]

Invasive Mechanical Ventilation is used in cases where non-invasive methods are insufficient to support respiratory functions. Mechanical ventilation provides controlled support for oxygenation and ventilation by delivering positive-pressure breaths through an endotracheal tube (ETT) to the infant [3]. While invasive ventilation is effective in managing several respiratory failures, it is associated with lung injury, air leaks and respiratory diseases. Therefore, carefully monitoring and optimising the ventilator is essential to minimise complications and promote successful weaning off the ventilator[4].

In many healthcare facilities, there is a shortage of physicians available to provide direct care to patients. This is particularly true when caring for premature babies, who require more observation to survive the first few days of life [5]. Therefore, the need for autonomous systems in healthcare has become crucial in evaluating and monitoring these neonates to prevent common respiratory diseases like Bronchopulmonary Dysplasia (BPD) [5].

### 2.2 Importance of Carbon Dioxide Monitoring in Neonatal Care:

It is important to continuously monitor carbon dioxide levels in the Neonatal Intensive Care Unit (NICU) due to the increased risk of respiratory and neurological complications for newborn infants [6]. In neonatal care, respiratory support is commonly provided through ventilation, significantly impacting both short-term and long-term outcomes in neonatal health [7]. While monitoring oxygen, pressure, and volume are standard practices used to reduce infant morbidities, continuous and non-invasive monitoring of CO2 remains an underemphasised aspect in the NICU.

Extremely premature babies often experience high or low levels of carbon dioxide (CO2) in their blood, known as hypercarbia or hypocarbia. The presence of hypercarbia and hypocarbia are frequently associated with other lung issues, such as Bronchopulmonary Dysplasia (BPD) [7]. Continuous monitoring of end-tidal CO2 (etCO2) plays a pivotal role in achieving stability, particularly during critical phases such as in delivery rooms and receiving respiratory support [8].

There is currently a lack of agreement among experts regarding the ideal partial pressure of carbon dioxide (pCO2) levels in newborns. Similarly, the medical community has not reached a agreement on the benefits of allowing higher levels of CO2 (permissive hypercapnia) in neonatal care. One systematic review [9] suggested that maintaining pCO2 levels within the range of 5–7 kPa may be considered safe for this patient population. However, there are several challenges in effectively monitoring CO2 levels in neonates, such as issues with the position and leakage of the endotracheal tube (ETT). These issues may affect the accuracy of measurements.

Continuous non-invasive monitoring methods, such as etCO2 and transcutaneous CO2 (TcCO2), offer promising avenues for monitoring CO2 levels in neonates. EtCO2 monitoring, which is widely used in the NICU, measures the maximum level of carbon dioxide at the end of expiration. However, its accuracy can be compromised by a rapid respiratory rate and the presence of significant lung disease. Conversely, TcCO2 is a measurement method that measures CO2 by diffusion through the tissue and uses a sensor to monitor CO2.  This non-invasive method may better reflect arterial CO2 levels, especially during the transport of neonates, as it is less affected by such mismatches, which occur in etCO2 monitoring, usually because of leakage from the ETT tube[10].

### 2.3 Closed-loop Monitoring in Neonatal Care

Closed-loop control (CLC) systems are a type of autonomous technology commonly used in everyday life but limited in critical healthcare [5]. These systems use feedback from one or several monitored variables to control the output state, therefore ensuring a more personalised approach to patient care.

Personalised respiratory support is when infant parameters such as gestational age, weight and underlying respiratory diseases are used to adapt a strategy to meet the needs of each infant. By personalising the approach, outcomes can be optimised, and the risks associated with respiratory support interventions can be minimised [11].

Implementing CLC systems in neonatal care offers several advantages, including reducing healthcare workloads, improving care quality, maintaining supplies, and shortening mechanical ventilation durations [5]. By automating aspects of respiratory support and CO2 management, CLC systems reduce the monitoring burdens, allowing practitioners to focus on critical patient care, thus improving efficiency and resource allocation.

An example of incorporating Closed-Loop Control (CLC) into healthcare systems is the IntelliVent-ASV technology. This new technology shows promise in enhancing ventilation strategies through advanced monitoring techniques and closed-loop ventilation. The system aims to achieve precise control of respiratory parameters while minimizing the risks of lung injury and adverse outcomes (Arnal et al., 2013).

### 2.4 Issues with Accuracy and Reliability of Predicting $CO_2$

Accurately predicting CO2 levels in neonates presents several challenges due to the dynamic nature of respiratory physiology and the influence of various factors on CO2 exchange.

The respiratory system of newborn babies varies significantly in terms of lung compliance and lung physiology [13] These variations can affect the levels of CO2 in their bodies, making it difficult to predict CO2 dynamics in individual neonates. Moreover, factors such as gestational age, postnatal age, and underlying medical conditions further complicate the prediction of CO2 levels.

It is important to note that there are technical limitations associated with CO2 monitoring devices, such as end-tidal CO2 (etCO2) monitors and transcutaneous CO2 (TcCO2) monitors, which can negatively affect the accuracy and reliability of CO2 predictions. Factors, including apparatus dead space, sampling time, response time and calibration errors, may contribute to inaccurate CO2 measurements and subsequent predictions. Therefore, advancements in sensor technology and signal processing algorithms are required to address these technical limitations and improve the reliability of CO2 predictions in neonates [14]

### 2.5 Machine Learning and Explainable AI (XAI) in Neonatal Care

Explainable AI (XAI) is a system designed to provide understandable and intelligent explanations to individuals who have limited knowledge about machine learning models and algorithms. XAI systems can explain how a model works based on the field, dataset, and expectations of the system. Fully interpretable models offer complete transparency and clear explanations for complex models[15]. Machine learning and explainable artificial interference are promising tools that can enhance neonatal care. They can help improve clinical decision-making, forecasts, and patient outcomes. Machine learning models can analyse vast amounts of clinical data, such as vital signs, laboratory results, and images. By identifying patterns, predicting outcomes, and recommending personalised treatments, these models can assist healthcare professionals in providing more accurate care to neonates. [16].

Recently, machine learning (ML) algorithms have been used in neonatal care to enhance mechanical ventilation and recognise potential risks of respiratory outcomes [17]. These algorithms use data obtained from health records, physiological monitoring devices, and genetic studies to create predictive models that help clinicians identify issues early on and take necessary action. Some advantages of using Machine learning in Neonatal care are that it improves sensitivity, is easily interpretable, is cost-effective, and can detect trends and patterns that practitioners may not pick up when monitoring ventilation.

Although machine learning has the potential to revolutionise neonatal care, there are several challenges and considerations that must be considered to ensure its safe and effective implementation. One of the primary concerns is the need for rigorous validation of ML models using a large and diverse patient population [18]. Another challenge is integrating ML algorithms into clinical workflows while ensuring their predictions are interpretable and trustworthy. Additionally, ethical considerations such as patient privacy and algorithm biases must be carefully considered to prevent potential harm and ensure fair treatment for all patients [19]

While there are certainly challenges to overcome, the combination of Machine Learning and Explainable Artificial Intelligence (XAI) has the potential to greatly benefit neonatal clinical healthcare providers. By utilizing these technologies, healthcare providers can enhance the quality and efficiency of care delivery, leading to improved outcomes for vulnerable newborns.

### 2.6 Current Technology and Past Research

Explainable artificial intelligence (XAI) is a promising area of study for predicting CO2 levels in neonates. Although no research has been done in the particular field of XAI in CO2 monitoring in neonatal care, XAI has already demonstrated potential in various diagnostic and monitoring tasks in the NICU. Recent studies have shown that XAI can be used to classify medical thermal images of neonates using class activation maps, proving that XAI models can significantly aid practitioners in diagnosis and treatment management [20]

In addition to XAI, innovative monitoring devices are also entering the market for neonatal care. For instance, wearable sensor platforms and IoT-based neonatal monitoring systems with small textiles are being used to allow for continuous, noninvasive monitoring of vital signs like heart rate, temperature, and oxygen saturation [21]. Integrating such technologies with XAI could further enhance their predictive accuracy and extend to CO2 monitoring applications in the future.

Furthermore, past technologies such as IntelliVent-ASV have shown the feasibility of fully closed-loop control ventilation in intensive care settings, including patients with active respiratory failure. Such automation technologies in healthcare, coupled with intelligent XAI systems, could transform how respiratory support is provided to neonates. While IntelliVent-ASV has shown to be safe for use in ICU patients with various lung conditions, the extension of these findings to specific populations, such as neonates, requires further investigation [12]

Although there have been significant advancements in using machine learning to predict CO2 levels in neonates, there is limited research solely focused on the development of explainable AI (XAI) for CO2 prediction in Neonatal Care. The shift from traditional monitoring methods to the promising field of XAI highlights the importance of models that are not only powerful but also transparent and interpretable. Such models are essential to enhance respiratory support in neonatal care.

# SECTION 2

This section of the report provides an overview of the methodology used in the study. It starts by discussing the time-series forecast method used to predict end-tidal carbon dioxide (etCO2) levels in neonates. Next, it explains how tree-based models were chosen for their ability to leverage sequential data, which is reflective of the dynamic nature of neonatal health. The study also incorporates deep learning models, which are known to improve performance with increased datasets. All of these models were selected for their ability to use data windows, which were necessary because of the sequential nature of the dataset provided.

# 3 METHODOLOGIES:

### 3.1 Time Series Forecasting

Time Forecasting is the process of predicting future outcomes based on historical data and domain-specific knowledge and is a critical component in many fields, such as healthcare. One of the primary tools used in forecasting is time series analysis, which involves studying a sequence of data points collected over time[22].

In neonatal healthcare, implementing a time series is essential as it captures the sequential progression of medical events, reflecting the inherently dynamic nature of a newborn's health conditions [13]. While time series analysis might initially appear similar to a standard regression problem, as historical data is used to predict a future value in regression, there are crucial differences as time series data has a specific order that is inherent to the data and must be preserved during modelling[22]. This order is a fundamental characteristic of time series data and distinguishes it from other types of data used in regression models, which do not possess this inherent order.

Any rearrangement of time series data could compromise the accuracy of predictions. This is particularly critical when predicting end-tidal carbon dioxide (etCO2) levels in newborns, where accurate forecasting can significantly impact clinical decisions and patient outcomes. Therefore, understanding the unique characteristics of time series data and the appropriate analytical techniques is essential for reliable and accurate predictions in healthcare analytics.

### 3.2 Forecasting with Tree-based Models

Tree-based models are becoming increasingly popular for forecasting trends over time, as they can handle complex patterns that don't follow a straight line. Gradient boosting is a standout technique in this field, and it is known for making very accurate predictions by learning from past mistakes.

Tree-based methods include decision trees, random forests, and gradient boosting. Decision trees split up the data into different groups, while random forests combine many trees to obtain more reliable results. Gradient boosting can be used for both regression and classification problems. It creates a predictive model by combining decision trees, which are weak predictive models. The Gradient boosting model is built in stages and is generalised to optimise a differentiable loss function (measures the difference between the predicted and actual values). Gradient boosting is a machine learning technique that involves creating a collection of decision trees, where each tree corrects the mistakes of the previous one. This is done by adding predictions sequentially, which results in an improved model performance. To measure the performance of the model, a loss function is utilised. The new models that are added to the existing ones predict the residuals or errors of the prior models. All these models are then combined to make a final prediction and work best [23]

XGBoost, which stands for Extreme Gradient Boosting, is an implementation of gradient boosting that is both powerful and efficient. It utilises a penalty, known as a regularisation term, to mitigate overfitting by penalising complex models. XGBoost is capable of efficiently handling data that is dispersed by selecting the best direction for missing values at each node in a tree. It is designed to improve system performance and run operations in parallel, making it highly effective in data analysis. XGBoost's success over other tree-based models is due to

The model achieves scalability and high performance through a combination of software and hardware optimization techniques. Its processing capability can be enhanced by adjusting the number of trees and the depth of each tree. These techniques enable the model to efficiently handle large datasets [24].

A study conducted by El Houssainy on using time series to forecast gold prices [25] demonstrated the powerful capabilities of Gradient Boost. Although Random Forest was found to be the most accurate method, Gradient Boosting performed exceptionally well due to its ability to capture nonlinear relationships and handle different types of data, such as financial information, without requiring a lot of preparation work. This makes it a robust option for forecasting, particularly when compared to traditional statistical methods like ARIMA (Autoregressive moving average) models.

### 3.2 Forecasting with Deep Learning Models

Deep learning is a technology that has branched from artificial neural networks (ANN). It involves creating models based on neural network architecture. The advantage of using deep learning is its ability to work with huge datasets and perform better as more data becomes available, especially for high-dimensional time series forecasting. In recent years, the field of deep learning has infiltrated several sectors such as robotics, natural languages, and medical fields[26].

Deep learning models are an excellent choice for developing time series models due to their high accuracy and their ability to easily incorporate data windowing. Data windowing is a crucial process in forecasting, which involves selecting a consecutive range of data points (known as a 'window') to forecast future events. This process works by defining the inputs and the corresponding labels (known outputs) and feeding them into the model for predicting future outputs[22].

There are two primary approaches for making predictions with data windowing. The first is a single-step approach, which predicts only the next future value. The deep learning model can accept any length of input, but it is limited to only one output for a single step. The second technique is the multiple-step approach, which takes in a sequence of specified inputs and predicts a sequence of future values, allowing for forecasting multiple time series ahead. The deep learning model, in this case, accepts any length model and outputs multiple predictions [22].

This study discusses five different deep learning models used in machine learning. It will start with the simplest deep learning model, linear regression [22], and then progress to more complex techniques such as DNN, CNN, LSTM and AR-LSTM

#### 3.2.1 Linear and Nonlinear Model:

A linear regression model is argued to be one of the simplest deep learning models to implement as it is just a neural network with no hidden layer. his model takes in all the features provided and multiplies weights to each respective feature, as shown in Figure 1. Next, the sum of the weighted features is calculated, and this sum is used as the prediction for the next time step.

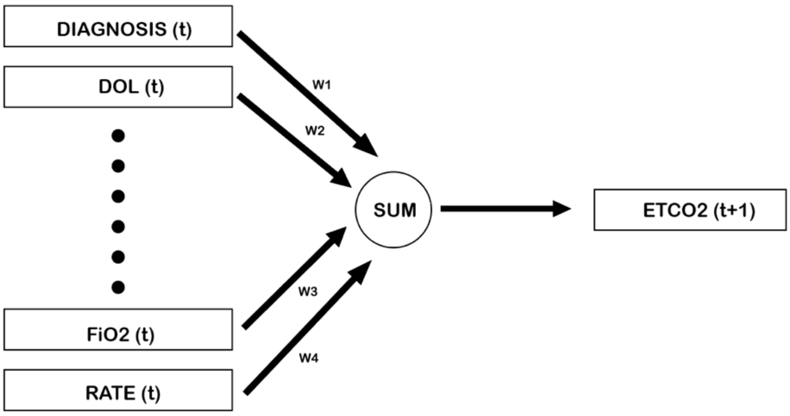


*Figure 1: Example of Standard Linear Model*

### 3.2.2 Deep Neural Networks (DNN):

A deep neural network is a type of model that consists of an input layer and an output layer, with several layers of neurons in between. The number of layers can vary depending on the complexity of the model. Nonlinear activation functions such as ReLU, sigmoid, and tanh are used in every neuron of the network to capture nonlinear relationships in the provided dataset. These functions generate output from the input data. Adding more hidden layers can improve the performance of the DNN, as it has more data to learn from. However, increasing the number of layers can also reduce the computational speed of the model [22].

### 3.2.3 Recurrent neural network (RNN):

Recurrent Neural Networks (RNNs) are a type of deep learning that is specifically designed to handle sequential data. RNNs work by incorporating a "hidden state," which is essentially an internal loop in the neural network architecture. This loop processes the output of the network by using past information as input. In practice, this feedback loop is essential to simulate the effect of "memory." However, a disadvantage of RNNs is that they can suffer from short-term memory issues due to the problem of vanishing gradients [22].

In machine learning, the problem of vanishing gradients arises when the gradient of the loss function becomes very small during backpropagation. Backpropagation is a training process for neural networks where the model learns from its errors by adjusting its weights to minimize the difference between the actual and predicted output. When the gradient change is small, the weights of the network do not update much, which means the network stops learning. This issue is particularly pronounced in RNNs, which use backpropagation through time (BPTT) for training sequences of data. BPTT is a technique for training neural networks on sequence data where the error is calculated and propagated backward through each time step of the sequence. Consequently, as you move further down the sequence, the information from an early stage becomes less relevant because the network stops learning [27].

### 3.2.4 Long short-term memory (LSTM):

Long-short-term memory (LSTM) is a type of recurrent neural network (RNN) that solves the issue of the vanishing gradient. It does so by introducing a "cell state" that retains information from earlier values in the sequence. This state allows past values to stay in the neural network for a longer time, enabling computations to be carried out over extended periods [22].

The LSTM consists of three gates: the forget gate, the input gate, and the output gate.

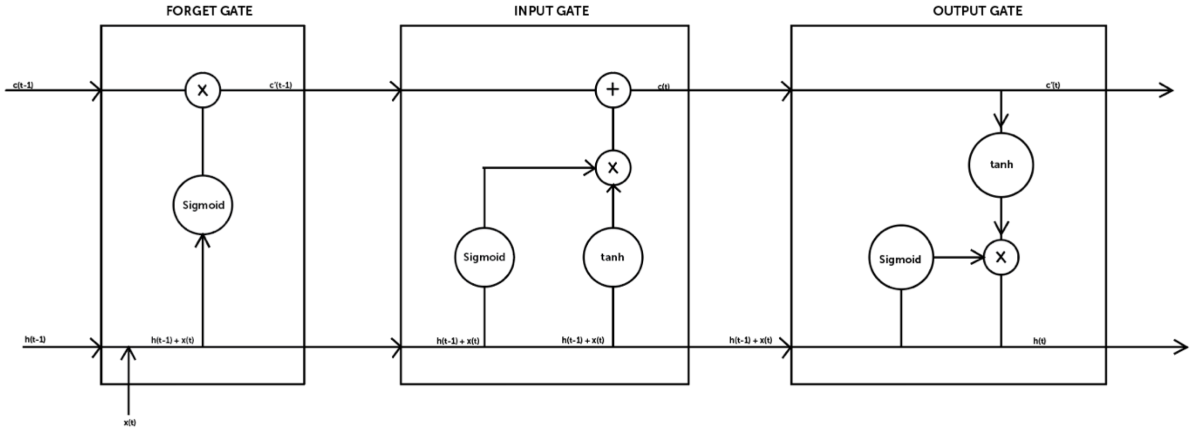


*Figure 2: Simple illustration of 'cell state in LSTM Neural Network* [22]

The forget gate evaluates whether the information from the previous steps is still relevant. In Figure 2, the previously hidden state h(t-1) and the current value of the sequence x(t) are both inputted into the forget gate. The h(t-1) and x(t) are then summed and duplicated. One copy is sent to the input gate, while the other copy is sent through a sigmoid function.

$$(i)\ f(t) = \frac{1}{1 - e^{-t}}$$

The sigmoid activation function helps decide which information to keep and which to discard. The value from the sigmoid is then used to create the new cell state c'(t-1) by multiplying it with the previous cell state c(t-1).

The input gate depicted in Figure 2 decides what information from the current step is relevant. The input gate receives a combination of the previous hidden state (h(t-1)) and the current element (x(t)) from the forget gate. This combination gets duplicated, with one copy directed towards the output gate and another copy sent through the sigmoid function. The third copy goes through the hyperbolic tangent (tanh) function which controls the input values before they are sent to the output gate. The output of the sigmoid and the tanh functions are combined by multiplication, and the result from this operation is then summed with the updated cell state from the forget gate c'(t-1). As a result, the final cell state c(t) is sent to the output cell.

The output gate decides what information should be sent to the next element of the sequence. This process uses the combination of the previous hidden state (h(t-1)) and the current element (x(t)) from the input gate and passes it through a sigmoid function. While the tanh function is current cell state c(t) from the input gate is passed through the tanh function. These resulting values from the sigmoid and tanh functions are then multiplied to generate a new hidden state, h(t). The current cell state c(t) is then sent from the output gate to the next neuron [22]. This process is illustrated in Figure 2 and is repeated multiple times within the neural network.

### 3.2.5 Convolutional neural network (CNN):

A Convolutional Neural Network is a type of deep learning model that uses convolution, a mathematical operation that applies filters called kernels to input data to extract essential features and patterns. These filters slide over the input data, performing element-wise multiplications followed by a sum, which produces a new feature map that highlights crucial information in the dataset, as seen in Figure 3. This process captures the hierarchy in data, usually images, as CNNs are commonly used for image recognition. CNN also help reduce the size of the data being processed, allowing the network to focus on the most relevant patterns and reducing the risk of memorising the training data too closely, which is known as overfitting. This allows for the addition of more convolution layers, enabling the network to process features thoroughly. However, to prevent rapid feature reduction, padding is applied. Padding involves adding zeros to both ends of the input vector in order to maintain the output dimensionality. In time series forecasting, CNNs usually employ one-dimensional (1D) convolution [22].

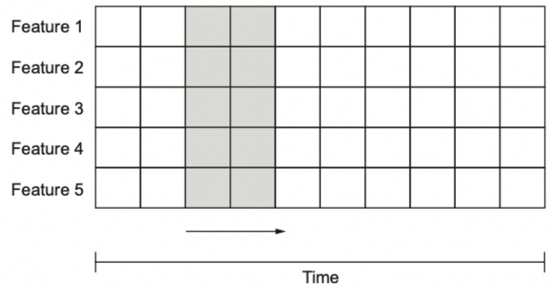


*Figure 3: Example of a kernel in CNN sliding across input data and performing element-wise multiplications* [22].

### 3.2.6 Autoregressive LSTM:

The Autoregressive Long Short-Term Memory (LSTM) network is a specialized form of LSTM models. Autoregressive models (AR) are models that predict future outcomes based on their previous predictions, forming a feedback loop. Therefore, Autoregressive LSTM is designed to remember and use past information from previous predictions to predict the next output. It works by taking the output from the previous time step as input for the prediction at the next time step. However, it is important to note that this method allows for error propagation because each prediction has some associated error. As a result, the longer the sequence grows, the more pronounced the error becomes [22].

# SECTION 3:

This section of the study is dedicated to the process of creating an XAI model that can accurately predict CO2 levels in neonates. The process began by collecting raw data from Prof. Theodore Dassios, Department of Women & Children's Health, King's College London. The collected data was then preprocessed and divided into three subsets - testing, training, and validation sets. Further data processing techniques were applied to refine the subsets to optimize the model's learning. A predictive model is then selected, and its hyperparameters are fine-tuned to improve its performance. The selected model was then trained on the training data to identify patterns and learn from the data. The model evaluation follows, where the validation set is used to assess the model's predictive accuracy. Finally, an explainable artificial intelligence analysis is conducted on the selected model to ensure that its predictions are transparent and understandable. This concludes the model development cycle. Figure 7 provides an overview of the steps involved in building an explainable AI model that predicts CO2 levels in neonates in the form of a flow diagram.

# 4 MODEL DEVELOPMENT:

### 4.1 Understanding the nature of the dataset and the predicted value of etCO2

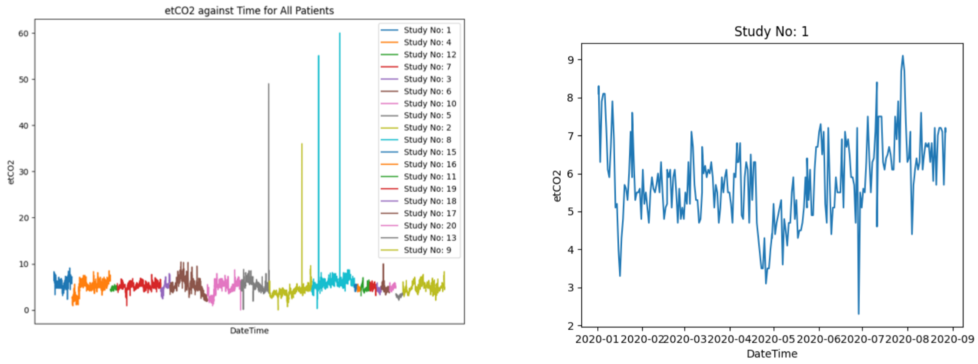


*Figure 4 a) and b): EtCO2 plot against time for all patients in the dataset, as well as a zoomed-in evaluation of patient 1 from Study 1.*

To begin analysing etCO2's time series data, it is helpful to break the data down into three components to better understand the patterns present. These three components are the trend, the seasonal, and the residual components. The trend represents a slow-moving change in a time series that can either increase or decrease over time. Seasonality points to a pattern in the data that repeats cyclically over time. Finally, the residual is the random error of the data points, which is often referred to as white noise.

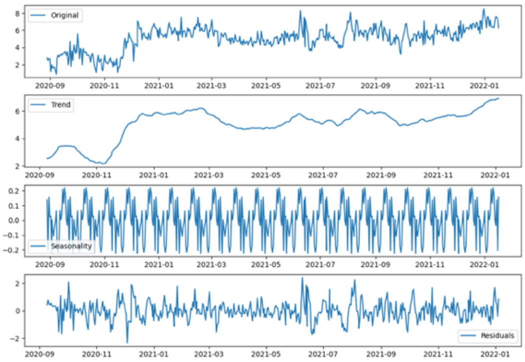


*Figure 5: A breakdown of etCO2 time series data into its three components.*

Figure 5 demonstrates that there is no significant trend in the etCO2 measurements. The measurements remain reasonably consistent without any noticeable increase or decrease across multiple datasets from different patients. While Figure 5 may indicate a repeating pattern that could be misconstrued as seasonality, this pattern is more closely linked to the natural breathing rhythms of the patients, as evidenced by the individual patient data in Figures 4a and b. To confirm the lack of inherent seasonality in the etCO2 data measurements and verify the stationary nature of the series, an Augmented Dickey-Fuller (ADF). ADF test was conducted test is a statistical test that checks if a time series data is non-stationary, which means it does not have a consistent mean, variance, and covariance over time. If the ADF test gives a p-value more than 0.05 and the test statistic is far from zero in the negative direction, then we can say the data is actually stable and consistent over time, not influenced by trends or seasonal effects. When applied to the etCO2 time series data, the ADF test showed a p-value over 0.05, indicating that the etCO2 data is stationary and that its fluctuations are likely due to short-term, random variations rather than long-term trends or predictable seasonal patterns.

Moreover, to gain a thorough understanding of the other features in the dataset, a histogram distribution analysed various numerical features in the dataset. These features are displayed on histograms to detect any anomalies or common patterns, such as normal or skewed distribution. The histograms provide insights into the following characteristics:

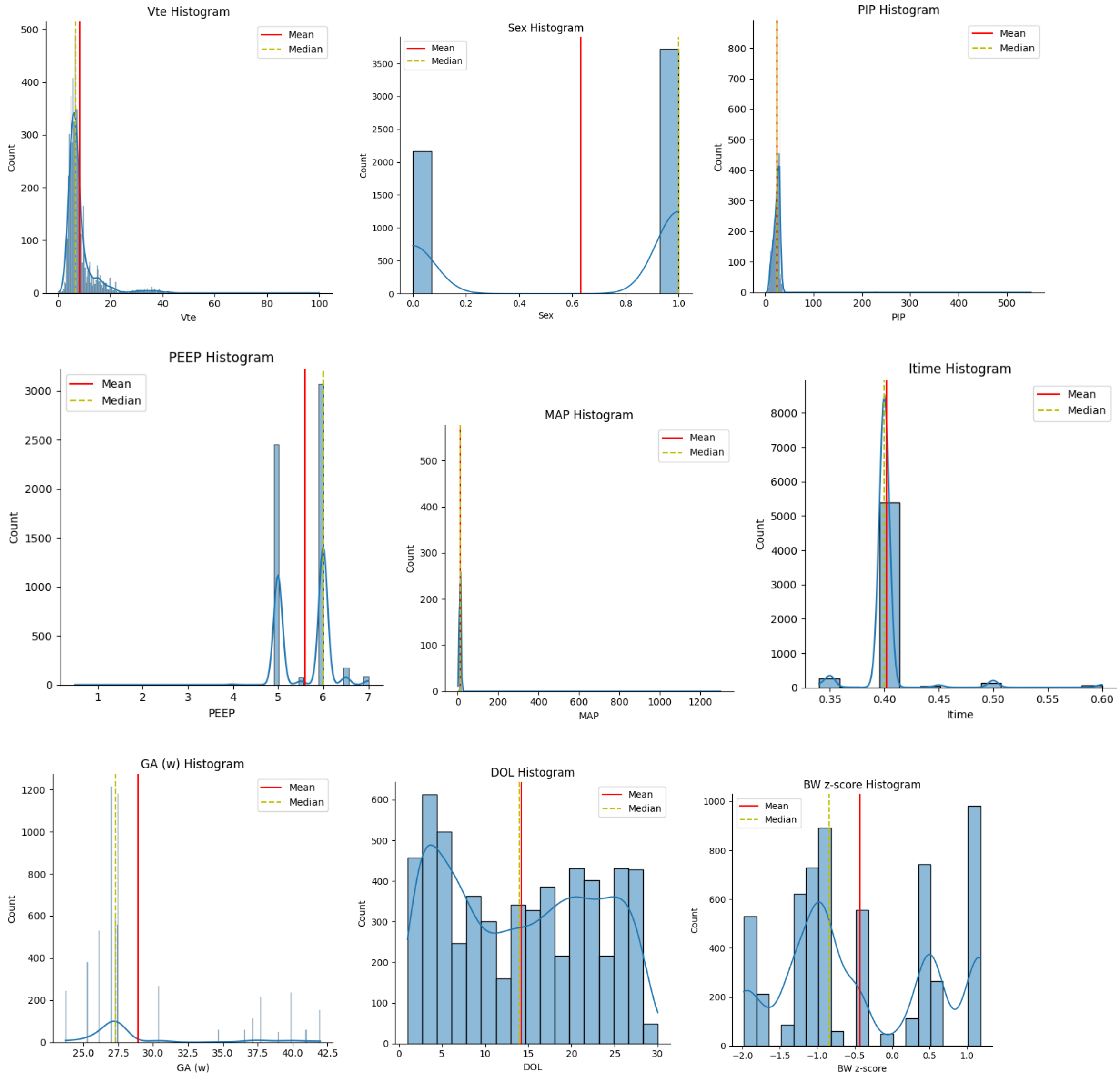


Figure 6a) to k): Histograms representing all the numerical features in the dataset Vte, Sex, PIP, PEEP, MAP, Itime, GA, DOL and BW z-score.

In Figures 6a to 6k, it is observed that the tidal volume (Vte) and peak inspiratory pressure (PIP) histograms exhibit right-skewed distributions, with the majority of data clustered at lower values and the mean higher than the median. In contrast, the positive end-expiratory pressure (PEEP) is left-skewed, with data leaning towards the higher end and a mean lower than the median. The sex distribution is distinctly bimodal, reflecting two separate gender groups, while the mean airway pressure (MAP) shows a sharp peak with data tightly centred around the mean, suggesting little variability. The inspiratory time (Itime) resembles the Vte in its right skewness, and the gestational age (GA) displays a multi-modal pattern, indicating multiple distinct subgroups. The duration of labour (DOL) presents a near-normal but slightly skewed distribution, as seen in the overlay of the histogram and density plot. Overall, these histograms illustrate varied skewness patterns, indicating irregular distributions within the dataset.

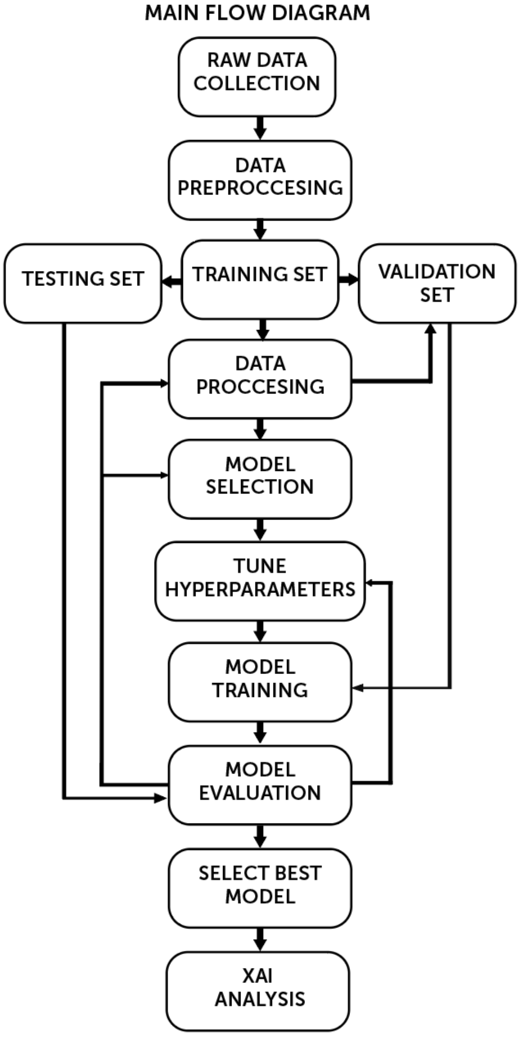


*Figure 7: Block diagram of the model development*

## 4.2 DATA PRE-PROCESSING

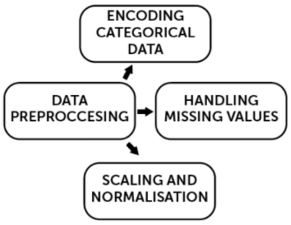


*Figure 8: Break down of Data-pre-processing Tasks*

Data pre-processing is a crucial step in preparing data for analysis. This process, as shown in Figure 8, involves handling missing data, data augmentation, encoding categorical data into numerical format, and scaling and normalising numerical data to ensure that they are on the same scale.

The dataset used for analysis consisted of 5894 instances and 22 features. Among these features, 10 were independent of time, while the other 12 were dependent on time. In order to ensure accurate prediction of etCO2, irrelevant features like 'Hospital ID' were eliminated. Furthermore, 'BW (g)' was excluded and 'BW z-score' was selected as it provides standardized measurements.

To capture the cyclic nature of time, the 'Time' column entries, were transformed using sine and cosine functions. This transformation helped capture the periodic nature of time in a 24-hour cycle, as seen in Figure 9.

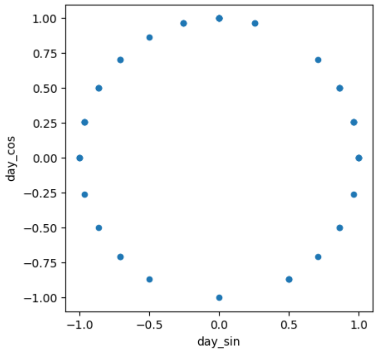


*Figure 9: Transformation of the Time column HH:MM to a 24-hour clock cycle*

### 4.21 Handling Missing Data

To address the missing data in the database, a function was used to calculate the total amount of missing data for each feature in the dataset. The results of this analysis revealed that the dataset was predominantly complete, with most of the features having a full set of data entries. Notably, a select few features exhibited instances of missing data, which required addressing to preserve the integrity of the data set and ensure the robustness of any subsequent analysis. The features 'Diagnosis' and ‘BPD’ had 60 missing entries. 'Ventilation Mode' was missing three entries, 'PIP' had 29, 'PEEP' was short by 20, and 'MAP' lacked 28 values. Additionally, the features 'Vte', 'Itime', 'Rate', 'Trigger Rate', 'Compliance', 'Resistance', 'FiO2', and 'etCO2' had missing values ranging from 21 to 383.

| **Study no** | **Hospital No** | **Sex** | **GA (w)** | **BW (g)** | **BW z-score** | **Diagnosis** | **BPD** | **DOL** | **Time** | **Ventilation mode** |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 60 | 60 | 60 | 0 | 0 | 3 |
| PIP | **PEEP** | **MAP** | **VTE** | **Itime** | **Rate** | **Trigger rate** | **Compliance** | **Resistance** | **FiO2** | **etCO2** |
| 29 | 20 | 28 | 42 | 21 | 23 | 383 | 35 | 37 | 25 | 334 |

*Table 1: Sum of the total missing values for each feature in the dataset.*

To handle missing data within the given dataset, a technique called targeted imputation was used. This approach involves substituting missing values with the median value from the same ‘Study no’, which is just the number associated with each patient. For example, if a patient's data from “Study no: 1” was incomplete, the gaps would be filled with the median value for each feature for that particular patient. This procedure was found to be highly effective due to the structured nature of the dataset, where patient information was classified sequentially. The median was chosen instead of the mean since most features exhibited a skewed distribution, as seen in Figures 6a)-k). The median is less sensitive to the influence of outliers, making it a more accurate representation of the central tendency for such data and, therefore, preserving the integrity of the data representation. Lastly, after all the missing data was handled, the ‘Study no’ feature was then dropped from the database as it held no value in predicting etCO2.

### 4.22 Encoding Categorical Data

To determine whether the data needed encoding, it was important to identify the data type in the data set and identify which features were categorical or numerical.

| **DATA TYPES:** | |
|---|---|
| **Sex** | int64 |
| **GA (w)** | float64 |
| **BW z-score** | float64 |
| **Diagnosis** | object |
| **BPD** | int64 |
| **DOL** | int64 |
| **Ventilation mode** | object |
| **PIP** | float64 |
| **PEEP** | float64 |
| **MAP** | float64 |
| **Vte** | float64 |
| **Itime** | float64 |
| **rate** | float64 |
| **trigger rate** | float64 |
| **compliance** | float64 |
| **resistance** | float64 |
| **FiO2** | int64 |
| **etCO2** | float64 |

*Table 2: The data type of each feature in the dataset*

After examining Table 2, it was observed that there are two categorical variables, namely "Diagnosis" and "Ventilation." To convert these non-numerical data types to numbers, one-hot encoding was used. This encoding technique was preferred over label encoding to avoid any possible misinterpretation of labels, which could lead to incorrect ranking of data by the model. However, one of the noticeable consequences of using one-hot encoding was an increase in feature dimensions. The feature increased from 18 to 37 after encoding. This increases the model's complexity and computational requirements.

### 4.23 Scaling and Normalising Numerical Data

The data within the database was both normalisation and scaled. This pre-processing step is necessary for Deep Neural Networks (DNNs), which use gradient-based optimisation algorithms like backpropagation to update weights during training. Without uniform feature scales, gradients could become imbalanced, potentially leading to slow convergence or divergence within the training process. Normalisation ensured that all features shared similar scales, thus enabling the optimisation algorithms to achieve more efficient convergence [22].

### 4.3 Data Splitting

The dataset has been divided into three subsets, namely the training set, the validation set, and the test set. This split follows the standard 70:20:10 ratio.

### 4.4 Feature Engineering

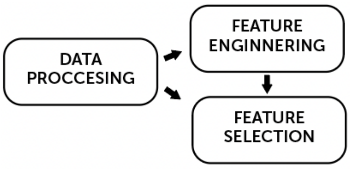


*Figure 10: Break down of Data Processing Tasks*

Through the process of feature engineering, a set of predictive features were formulated, some of which are inspired by existing research, while others were conceived by exploring various inter-feature relationships to determine their predictive capabilities.

The first feature engineered was 'Age_at_Diagnosis’, which calculated the age at which a neonate was diagnosed with respiratory diseases based on gestational age and DOL. This aligned with studies on age-related disease patterns in neonates[28]. 'Ventilation_Efficiency' was another feature that was engineered, and it was derived from tidal volume and pressure differentials (PEEP-PIP), with the intention of providing valuable insights into the effectiveness of ventilation strategies, which is essential for optimising respiratory outcomes [29]. 'Respiration_Health' feature aimed to quantify the balance between compliance and resistance in the respiratory system. The 'Oxygen_to_etCO2' feature aimed to assess oxygenation relative to end-tidal CO2 levels, highlighting the importance of adequate gas exchange in neonatal ventilation. The 'Is_low_birth_rate' feature was included to flag neonates with low birth weight, as a low birth rate poses a significant risk that is supported by research linking it to respiratory morbidity and mortality[30]. The 'Ventilation_Duration' feature aimed to assess the impact on neonates with prolonged periods of ventilation [31]. The 'Ventilation_and_BPD' binary feature aimed to identify ventilation-associated bronchopulmonary dysplasia (BDP) cases, aiding in understanding potential causal relationships between ventilation and disease. Finally, the 'Risk_Factor' feature integrated birth weight, presence of BDP, and gender with the aim of identifying infants with higher morbidity and mortality risk [31].

During the feature engineering, a minor subset of the data had positive and negative infinite values, which required their exclusion from the dataset, which was subsequently reduced from 5894x44 to 5861x44.

### 4.41 Feature selection

To enhance the predictive model's efficacy for neonatal care, features were selected to produce reliable forecasts with minimal error. This involved thoroughly testing different feature sets against the evaluation metric to find the most predictive model.

Two analytical techniques were used to identify the most significant features in this study. The first technique, SelectKBest from the Scikit-learn library, analysed univariate correlations to determine how individual features were related to the predicted value, etCO2. The second technique, the RandomForestRegressor, also from Scikit-learn, assessed the multivariate importance of features to identify those with strong predictive capabilities without causing model overfitting.

Different sets of features were created to train the models. The first set included all available features and served as a baseline. The second set was refined by selecting the top features identified by SelectKBest.

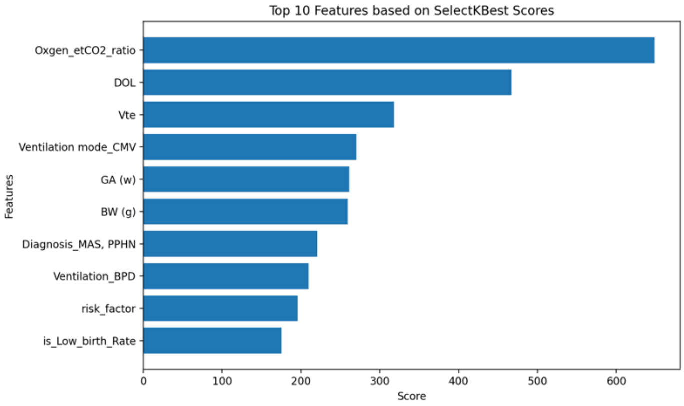


*Figure 11: Top 10 Features based on the SelectKBest Scores.*

Finally, the last set selected the features with the highest importance from Figure 12. These features were selected for their strong impact on the model for predicting etCO2.

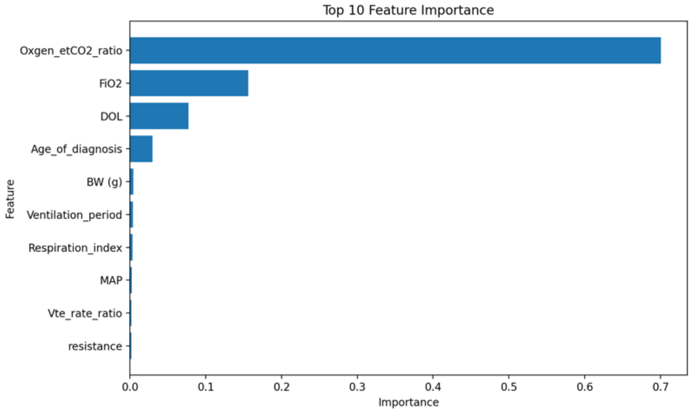


*Figure 12: Top 10 Feature Importance from the RandomForestRegressor.*

### 4.5 Data Augmentation

To ensure the optimal learning and training of the selected models, techniques were implemented to increase the size of the dataset provided, which was only 5894. The method used to augment the dataset was Feature noising. Feature noise is a straightforward method that involves adding a small amount of random noise to the existing features to generate new data points. As a starting point, a noise level of 0.02 was used as it was small enough to avoid drastically altering the data distribution but significant enough to improve model robustness. The size of the dataset increased to 11,720 instances.

### 4.6 Model Selection and Training

Six different models were chosen to create predictive models that can estimate the etCO2 in neonates. The models that were selected include linear regression, DNN, LSTM, CNN, XGBoost, and Autoregression LSTM. All deep learning models followed a similar implementation and evaluation training process for time series prediction using TensorFlow's Keras API. The tree-based model - XGBoost - used a technique known as k-fold cross-validation to train this model. All of these models were trained to handle different window sizes, ranging from 1 to 24 hours. Single-step prediction forecasts the CO2 level at the next time point, while multi-step prediction anticipates the CO2 levels for the upcoming 24 hours.

The accuracy of these predictions was evaluated, and the performance metrics were recorded for both the validation and testing stages. Mean Absolute Percentage Error (MAPE) is an evaluation metric that calculates the average difference between the predicted values and the actual values and then expresses this difference as a percentage of the actual values [32].

$$ii)\ MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{x_t - y_t}{x_t}\right| \times\ 100\%$$

Where n is the number of forecasts, $x_t$is the actual value for the at time t and $y_t$ is the predicted value for the at time t.

MAPE was selected as the appropriate evaluation metric because it provides an intuitive percentage-based measure that clearly communicates the average accuracy of forecasts relative to actual values, making it easily understandable and applicable across different data scales.

A new function called 'compile_fit' was developed to train the model with an early stopping feature. This function keeps track of the validation loss and stops the training when there is no significant improvement to avoid overfitting. This function compiled the model using the Adam optimiser, MAPE evaluation metrics, and Mean Squared Error (MSE) loss function. MSE measures the average squared difference between predicted and actual CO2 values [32]calculated using the following formula:

$$MSE = \frac{1}{n}\sum_{t=1}^{n}\frac{(y_t - y_t^*)^2}{n}$$

The function ‘compile_fit' also included a training process that runs for a certain number of epochs. An epoch represents a complete pass through the entire training dataset. The maximum number of epochs that can be set is 50. This means that the learning algorithm will pass through the data a maximum of 50 times. During each epoch, a subset of the dataset called a batch, is used for the training iteration.

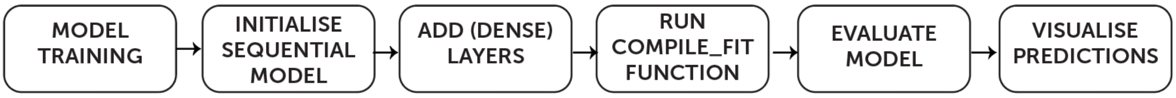


*Figure 13: Simple Breakdown of the Model Training for Deep Learning*

### 4.61 Linear Regression

The linear model was chosen as the baseline model because it is the simplest and most easily interpretable model. It was selected for its straightforward design and ease of comprehension. This model comprises of only two layers: an input layer that processes the data and an output layer that generates the prediction for CO2 levels. The output layer has a single neuron, emphasising the model's focus on predicting one variable at a time.

Linear regression models were created using TensorFlow’s Keras API that began with initialising a ‘Sequential’ model to arrange layers linearly. A single ‘Dense’ layer was added, tailored for single or multi-step forecasting.

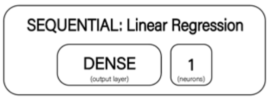


*Figure 14: Simplified representation of the layers in the Linear Regression model*

### 4.62 XGboost Model:

The XGBoost model was trained with specific parameters to ensure effective learning without overfitting. A learning rate of 0.01 was chosen for gradual learning, and the tree depth was capped at three to capture data patterns while maintaining simplicity. The model aimed to construct up to 1000 trees, with an early stopping set at 50 rounds to prevent unnecessary computations.

The method used to train the XGboost model was K-fold cross-validation:

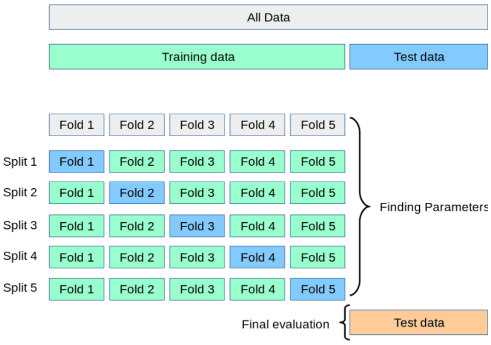


*Figure 15: Example of k-fold cross-validation* [33]

K-fold cross-validation is a technique used in machine learning to evaluate the performance of a model. Typically, this process involves dividing the dataset into two parts - the training and testing datasets. However, to refine the model's accuracy, the dataset was split into three groups: training, validation and testing. The training dataset was used to develop and fine-tune the model, while the validation and testing sets serve to evaluate its predictive ability and guard against overfitting, which occurs when a model learns the training data too well and fails to predict new, unseen data accurately.

To optimise the model, the training data was further split into 'k' number of folds. For example, in Figure 15, there are five-folds, which indicates 5-fold cross-validation. Through experimentation, it was found that using three folds was the best compromise between computational efficiency and low error rates. In this setup, one-fold is retained as the validation data for testing the model, and the remaining two (k-1) are used for further training. After each training and validation cycle, a performance metric (MAPE) is calculated. The benefit of using k-fold cross-validation with XGBoost is that it provides a robust estimate of the model's performance on unseen data. The training process involves repeatedly adjusting the model using different chunks of data, checking each time how well it's doing by comparing its guesses to actual CO2 levels.

Building more trees can make the model more accurate, but too many might make it too complex and not work well on new data. Adjusting how much change each tree makes can help the model learn better over time.

### 4.63 Deep Neural Networks

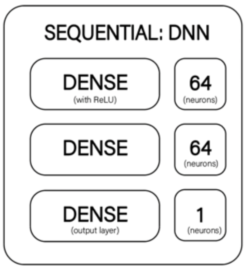


*Figure 16: Simplified representation of the layers in a DNN*

When training Deep Neural Network (DNN) models, a Sequential model structure was used. This structure consists of two hidden Dense layers, each containing 64 neurons. The choice of using 64 neurons in a DNN layer strikes a balanced compromise between the network's ability to learn complex patterns in the data and maintaining manageable computational demands. The activation function used was Rectified Linear Unit (ReLU). This function enables non-linear processing and enhances the model's ability to identify non-linear relationships within the data. Finally, the output 'Dense layer' produces a single predictive output.

### 4.64 Long Short-Term Memory

When training the LSTM model, a Sequential model was used, which included an LSTM layer with 64 neurons connected to a Dense output layer consisting of a single neuron. The LSTM layer's function was to process the input data sequentially and capture time-series dependencies. Following this layer, a Dense layer with a single neuron was included, serving as the output to provide the final prediction. Additionally, the Dense layer was included to simplify the LSTM's complex outputs into a single actionable output.

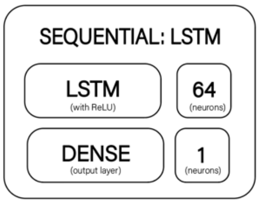


*Figure 17: Simplified representation of the layers in a LSTM*

### 4.65 Convolution Neural Networks

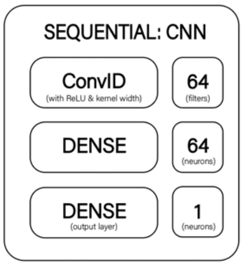


*Figure 18: Simplified representation of the layers in a CNN*

The CNN was trained using a Sequential model that comprises Conv1D and Dense layers. The CNN begins with a Conv1D layer with 64 filters of width 3 (kernel width). This layer convolves the input features. The data then passes through two Dense layers. The first Dense layer has 64 units to transform the features further, followed by a final single-unit Dense layer that outputs the predicted etCO2 value for the next time step.

To train the CNN model for single-step prediction, the kernel size was three timesteps wide. This means that the model requires an input sequence that spans three timesteps to make a single prediction. To predict 24 future values without interruption, the input sequence must be elongated to 26 timesteps (24 + KERNEL_WIDTH - 1). This ensures that no padding is needed.

### 4.66 Autoregression LSTM (AR-LSTM)

AR-LSTM was trained using a Sequential model that comprises three layers: ‘Dense’, ‘RNN’, and ‘LSTMCell’. The ‘LSTMCell’ is a detailed layer that provides access to state information and predictions. The ‘RNN’ layer trains the ‘LSTMCell’ with input data, and its output is then transformed into a prediction by the ‘Dense’ layer.

## SECTION 4:

This section will discuss the results obtained from developing six different models. To ensure the robustness of each model, four tests were conducted on each model. These tests helped to improve the performance of the system and identify the best-performing model. The four tests involved assessing the impact of feature engineering on the dataset, the effects of increasing the dataset through feature nosing, testing how the performance of the model changes when the prediction window is varied, and finally, testing how feature selection affects the model. By conducting these tests, the most ideal parameters of the models were selected to obtain the best-performing model.

# 5 RESULT:

## 5.1 Assessing the Impact of Feature Engineering on the Dataset:

A test was conducted to evaluate the overall impact of feature engineering on the model's performance. Instead of analysing the impact of each individual feature, all features were considered at once. Later, feature selection and importance analysis were performed to determine which features from the entire dataset had the most significant impact on and improved the model's performance.

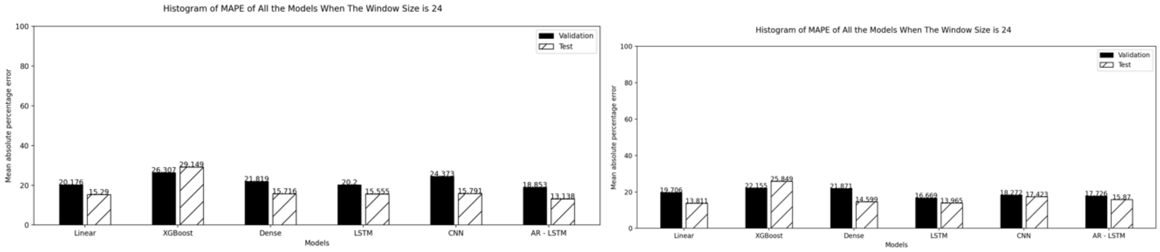


*Figure 19a) and b): Histogram of MAPE of all models without engineered features and with the engineered features, respectively:*

The aim of this test was to assess whether adding the features obtained from feature engineering would improve the model's performance. As seen in figures 19a) and b), Linear regression, XGBoost, DNN, and CNN performed better with the inclusion of the new features, while LSTM and AR-LSTM showed only a slight improvement. Overall, the addition of the new features had a positive impact on the model's performance. In subsequent tests, such as feature selection and feature importance, further investigations into the most impactful features will be discussed.

## 5.2 Assessing the effects of increasing the Dataset using Feature Noising

The dataset was augmented with varying noise percentages from 0% to 5%, as described in 4.15. After feature noising the whole dataset with different noise percentages, the dataset size doubled from 5893x44 to 11721x48.

These tests were carried out using the 24-hour window for all the trained models, and the MAPE results for each model are recorded in Table 3.

| | Linear regression | | XGBoost | | DNN | | LSTM | | CNN | | AR-LSTM | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Noise percentage | Validation | Test | Validation | Test | Validation | Test | Validation | Test | Validation | Test | Validation | Test |
| 0% | 19.7 | 13.8 | 22.2 | 25.8 | 21.9 | 14.6 | 16.7 | 13.9 | 18.2 | 17.4 | 17.7 | 15.9 |
| 1% | 35.0 | 16.8 | 6.6 | 5.1 | 29.4 | 11.9 | 32.4 | 15.1 | 29.8 | 14.2 | 28.6 | 13.7 |
| 2% | 26.2 | 16.3 | 4.8 | 5.4 | 19.8 | 12.3 | 23.5 | 15.5 | 21.3 | 13.7 | 22.8 | 18.0 |
| 3% | 31.8 | 17.7 | 6.6 | 6.3 | 24.9 | 12.5 | 27.6 | 15.0 | 25.6 | 13.8 | 27.1 | 14.8 |
| 4% | 27.4 | 17.4 | 12.1 | 11.4 | 21.2 | 13.3 | 24.8 | 15.4 | 20.8 | 13.5 | 22.1 | 13.8 |
| 5% | 40.3 | 18.3 | 10.7 | 7.7 | 38.8 | 14.9 | 38.5 | 17.6 | 36.8 | 14.4 | 32.5 | 14.0 |

*Test 3: Varying the noise percentage across all models and recording the corresponding MAPE for each model.*

Table 3 provides a comparative analysis of all the models used in this report. The dataset was tested with varying levels of noise, from 0% to 5%, to assess each model's performance. The introduction of a small increase in noise by 1% considerably improved XGBoost's performance, with its MAPE score dropping from 22.2% to 6.6% for the validation set. This suggests that XGBoost responds well to a larger training set and is less sensitive to noise in the training data. However, the other models did not respond to this increase in dataset size and performed worse.

As the noise levels increased, the XGBoost model remained robust, producing the lowest MAPE at 2% noise. DNN started performing better at 2% noising, with the MAPE decreasing from 21.9% to 19.8% for validation. However, every model's performance started to deteriorate beyond 3% noise, making testing and validation worse. This implies that the threshold of adding noise has been reached, and any further increase in noise will negatively impact model accuracy.

The analysis revealed that augmenting the dataset by introducing noise only leads to improved performance of XGBoost and DNN. XGBoost exhibits significant tolerance to high levels of noise. After conducting experiments with different noise levels, 2% noise was selected as the optimal percentage to use for increasing the dataset size due to XGBoost's exceptional increase in performance. Therefore, this percentage of noise was chosen as the optimal amount to increase the dataset size.

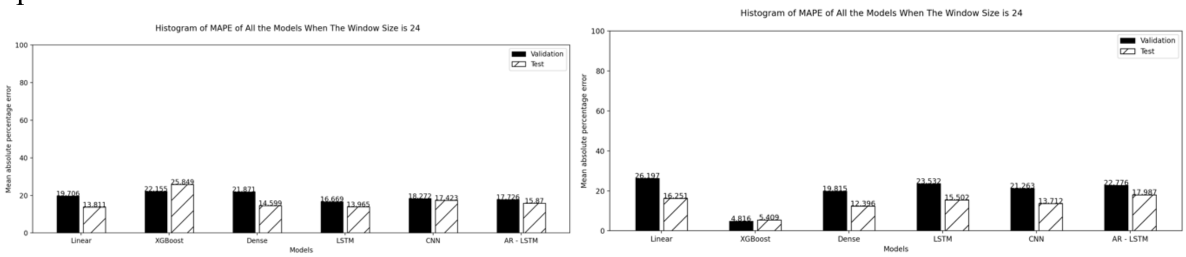


*Figure 19 a) and b) show the MAPE for all the models for Window Size 24, with 0% and 2% feature noising, respectively.*

### 5.3 Assessing the effects of varying the size of the window:

This test aimed to determine the optimal window size for predicting etCO2 by testing windows ranging from 1 to 24 hours and comparing their performance. This range was selected because the data was captured hourly over several days.

| | Linear regression | | XGBoost | | DNN | | LSTM | | CNN | | AR-LSTM | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MAPE | Validation | Test | Validation | Test | Validation | Test | Validation | Test | Validation | Test | Validation | Test |
| 1 | 25.5 | 13.1 | 30.9 | 28.2 | 17.9 | 7.7 | 17.0 | 8.0 | 18.5 | 9.0 | 18.0 | 7.3 |
| 4 | 24.9 | 14.9 | 13.4 | 12.4 | 22.4 | 13.7 | 20.4 | 13.0 | 18.5 | 11.6 | 15.0 | 8.3 |
| 8 | 23.4 | 15.0 | 7.3 | 7.3 | 19.4 | 12.4 | 19.7 | 13.2 | 19.7 | 13.5 | 14.7 | 9.7 |
| 12 | 24.7 | 16.5 | 6.6 | 7.3 | 19.1 | 13.1 | 20.6 | 14.7 | 19.3 | 12.6 | 14.5 | 9.8 |
| 16 | 27.9 | 16.8 | 10.6 | 9.6 | 23.3 | 12.8 | 24.9 | 14.9 | 22.1 | 12.4 | 18.8 | 13.1 |
| 20 | 24.9 | 16.8 | 7.5 | 8.3 | 22.5 | 15.0 | 22.0 | 15.0 | 18.3 | 13.2 | 27.1 | 19.1 |
| 24 | 26.2 | 16.3 | 4.8 | 5.4 | 19.8 | 12.3 | 23.5 | 15.5 | 21.3 | 13.7 | 22.8 | 18.0 |

*Table 4: Varying the window size across all models and recording the corresponding MAPE for each model.*

Based on the data presented in Table 4, it can be concluded that the Linear Regression and CNN models perform consistently well across different window sizes, ranging from 1 to 24 hours. However, it is worth noting that XGBoost outperforms all other models when the window size was set to 24 hours, with a remarkably low MAPE value of 4.8% for validation. On the other hand, XGBoost performs poorly when the window size is set to 1 hour, with a much higher MAPE value of 30.9%. This suggests that XGBoost works best when a larger number of past inputs are used to predict the next outcomes.

As the time window size increases, the performance of LSTM, DNN and AR-LSTM models in terms of MAPE also increases. Additionally, the AR-LSTM model's performance deteriorates when the time window is set to 20 hours, with the MAPE in the test set increasing to 27.1% from 18% at a window size of 1 hour. These observations suggest that these models struggle when a larger time window is used.

CNN's showed small variations across different window sizes, suggesting some robustness. The variations in CNN's MAPE scores suggest that the model could use larger windows without significantly decreasing performance.

After analysing the results, it is evident that the window size of 24 hours resulted in the best accuracy for XGBoost model. Therefore, this window size was selected as the optimal size for testing and predicting etCO2 values.

### 5.4 Assessing the effects of varying the number of features with different feature sets used:

Six predictive models were tested using three different feature combinations. The first set, referred to as Set 1, was the controlled set that included all the features in the dataset. The features in Sets 2 and 3 were selected based on the results of the top features obtained from the Selectktest and RandomTreeForest methods, respectively. These features were then incrementally increased from 1 to 10 to identify the smallest MAPE and determine the most accurate predictor model that performed the best.

**Feature Set 1:**

Feature Set 1 included all 44 features in the dataset with a prediction window 24.

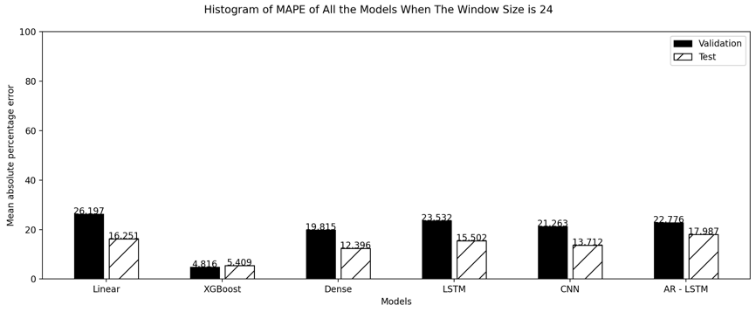


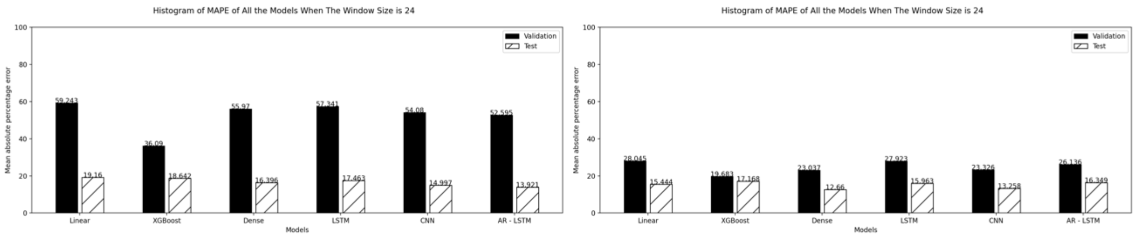
Figure 20: Histogram of all the MAPE Models using all the features in the dataset with a window 24.

**Feature Selection Set 2:**

For Feature Set 2, the top 10 features from SelectKBest were incrementally increased from 1 to 10 to find the smallest MAPE.

| | Linear regression | | XGBoost | | DNN | | LSTM | | CNN | | AR-LSTM | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MAPE | Validation | Test | Validation | Test | Validation | Test | Validation | Test | Validation | Test | Validation | Test |
| Oxygen _etCO2_ratio | 59.2 | 19.1 | 36.0 | 18.6 | 55.9 | 16.3 | 57.3 | 17.4 | 54.0 | 14.9 | 52.5 | 13.9 |
| DOL | 35.1 | 19.2 | 24.2 | 17.8 | 30.3 | 15.6 | 30.4 | 15.5 | 30.1 | 15.6 | 30.0 | 15.0 |
| Vte | 28.6 | 15.9 | 21.5 | 18.9 | 27.4 | 16.6 | 28.1 | 17.1 | 23.3 | 14.4 | 24.7 | 15.3 |
| Ventilation mode_CMV | 46.0 | 16.3 | 29.9 | 19.0 | 43.2 | 15.5 | 42.0 | 15.7 | 39.2 | 14.3 | 37.9 | 13.4 |
| GA(w) | 34.9 | 15.5 | 23.3 | 18.3 | 32.7 | 14.3 | 31.5 | 15.2 | 34.8 | 15.4 | 30.2 | 14.9 |
| BW(g) | 80.0 | 15.0 | 49.2 | 16.7 | 82.4 | 13.2 | 78.7 | 15.3 | 80.0 | 13.0 | 72.8 | 13.5 |
| Diagnois _MAS_PPHN | 40.7 | 14.3 | 25.8 | 16.7 | 37.8 | 12.1 | 38.7 | 15.1 | 37.8 | 15.1 | 35.0 | 13.4 |
| Ventilation BPD | 29.45 | 17.6 | 19.2 | 16.7 | 23.5 | 13.5 | 25.4 | 14.3 | 25.9 | 17.1 | 22.4 | 12.6 |
| Risk_factor | 41.9 | 16.2 | 25.8 | 16.7 | 38.9 | 13.8 | 41.1 | 17.4 | 37.0 | 13.6 | 34.7 | 13.3 |
| Is_lower _birth_rate | 28.0 | 15.4 | 19.7 | 17.1 | 23.0 | 12.7 | 27.9 | 15.9 | 23.3 | 13.3 | 26.1 | 16.3 |

*Table 5: MAPE results for each model when the number of features is from 1 to 10 from the SelectKBest test.*



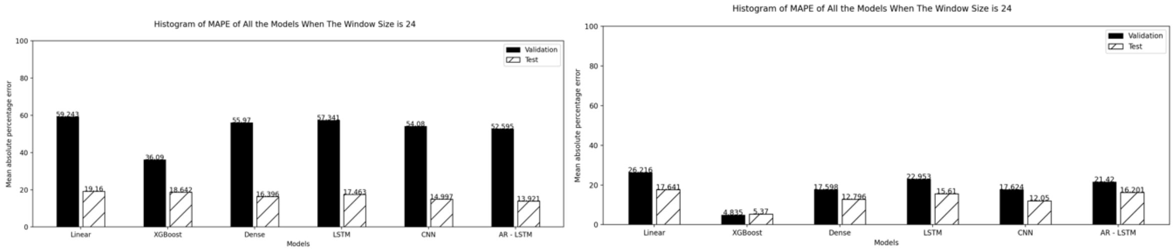
*Figures 21 and 22: Histogram of all the MAPE Models using only 1 feature and all 10 features, respectively, in the dataset with a window 24.*

**Feature Selection Set 3:**

For Feature Set 3, the top 10 features from RandomTreeRegression were incrementally increased from 1 to 10 to find the smallest MAPE.

| | Linear regression | | XGBoost | | DNN | | LSTM | | CNN | | AR-LSTM | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MAPE | Validation | Test | Validation | Test | Validation | Test | Validation | Test | Validation | Test | Validation | Test |
| Oxygen etCO2_rate | 59.2 | 19.1 | 36.0 | 18.6 | 55.9 | 16.3 | 57.3 | 17.4 | 54.0 | 14.9 | 52.5 | 13.9 |
| FiO2 | 51.8 | 19.8 | 12.0 | 4.2 | 51.3 | 17.1 | 50.9 | 17.4 | 48.7 | 15.4 | 45.2 | 13.5 |
| DOL | 32.4 | 18.0 | 6.9 | 5.4 | 29.6 | 16.1 | 32.3 | 17.7 | 28.6 | 15.0 | 31.5 | 17.3 |
| Age_of_ Diagnois | 35.2 | 17.6 | 10.5 | 7.4 | 34.3 | 17.4 | 32.9 | 16.5 | 32.8 | 17.8 | 33.3 | 17.2 |
| BW | 29.3 | 17.1 | 8.1 | 7.5 | 21.7 | 12.7 | 25.8 | 15.2 | 22.4 | 12.7 | 26.4 | 15.9 |
| Ventilation_period | 26.2 | 17.6 | 4.8 | 5.4 | 17.5 | 12.7 | 22.9 | 15.6 | 17.6 | 12.1 | 21.4 | 16.2 |
| Respiration_index | 42.6 | 17.1 | 9.1 | 5.0 | 36.7 | 11.8 | 38.0 | 11.8 | 38.0 | 13.5 | 39.7 | 16.3 |
| MAP | 27.0 | 17.3 | 5.4 | 5.2 | 21.0 | 11.5 | 26.2 | 15.0 | 23.9 | 15.0 | 24.4 | 14.9 |
| Vte_rate ratio | 29.1 | 16.4 | 7.5 | 6.7 | 20.8 | 12.2 | 25.9 | 15.0 | 23.6 | 14.3 | 22.1 | 13.3 |
| Resistance | 32.8 | 17.8 | 11.0 | 8.5 | 23.7 | 12.2 | 29.5 | 17.5 | 26.1 | 13.4 | 24.6 | 13.4 |

*Table 5: MAPE results for each model when the number of features is from 1 to 10 from the RandomTreeRegressor test.*



*Figures 23 and 24: Histogram of all the MAPE Models using only 1 feature and 6 top features, respectively, in the dataset with a window 24.*

After evaluating three different feature sets, it was observed that Feature Set 3, which was in close competition with Feature Set 1 (containing all features in the dataset), resulted in all models having the lowest MAPE. Although Feature Set 1 performed well, it wasn't considered the most desirable because of its complexity and lack of interpretability.

Feature Set 2 showed better performance as the number of features increased from 1 to 10. The highest performance was achieved at feature 10, which was 'Resistance'. However, its performance was not as high as Feature Set 3. Therefore, Feature Set 3, which contains six features including Oxygen_etCO2_ratio, FiO2, DOL, Age_of_diagnosis, BW, and Ventilation_period, was used to achieve the best performance across all models. It’s important to notes that the performance of all models decreased after Ventilation_period in Feature Set 3, resulting in a higher MAPE at Feature 10. Despite this, it was observed that Feature Set 3 overall performed better than Feature Set 2 and Feature Set 1.

Based on these results, Feature Set 3 was selected, and its top six features were used to predict etCO2. This selection combination of features resulted in the best-performing models with low MAPE values. Compared to the other feature sets, it was also more interpretable and less complex.

### 5.5 Analysing the overall performance of the models

After improving the dataset by applying feature engineering techniques, expanding its size using additional noise, and identifying the best set of features for model training, it is time to evaluate the MAPE of all the models. The final parameters selected for evaluating the model were a window size of 24, a feature noise percentage of 2%, and six selected features: Oxygen_etCO2_ratio, FiO2, DOL, Age_of_diagnosis, BW, and Ventilation_period. First, the MAPE will be used to compare each model's performance with the baseline model, which is a linear regression model in this case. This comparison will help determine if the other models have provided an improvement over the baseline. Then, all the models will be compared, and the best-performing model will be selected. The final step involves assessing the MAPE within a medical framework to confirm the model's accuracy is acceptable for practical use in a healthcare setting.

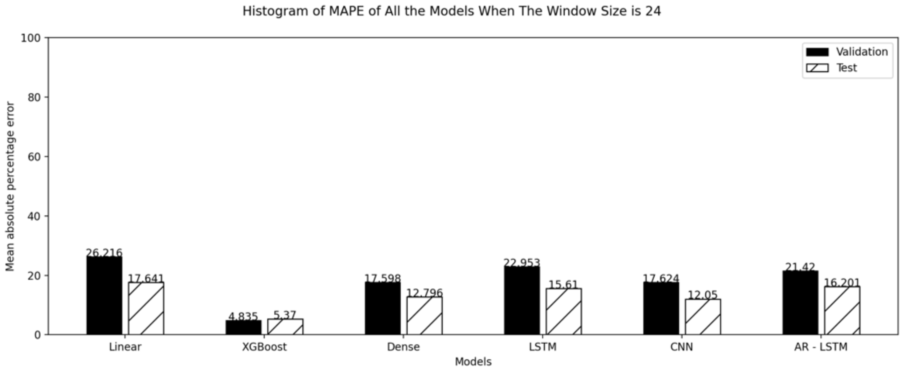
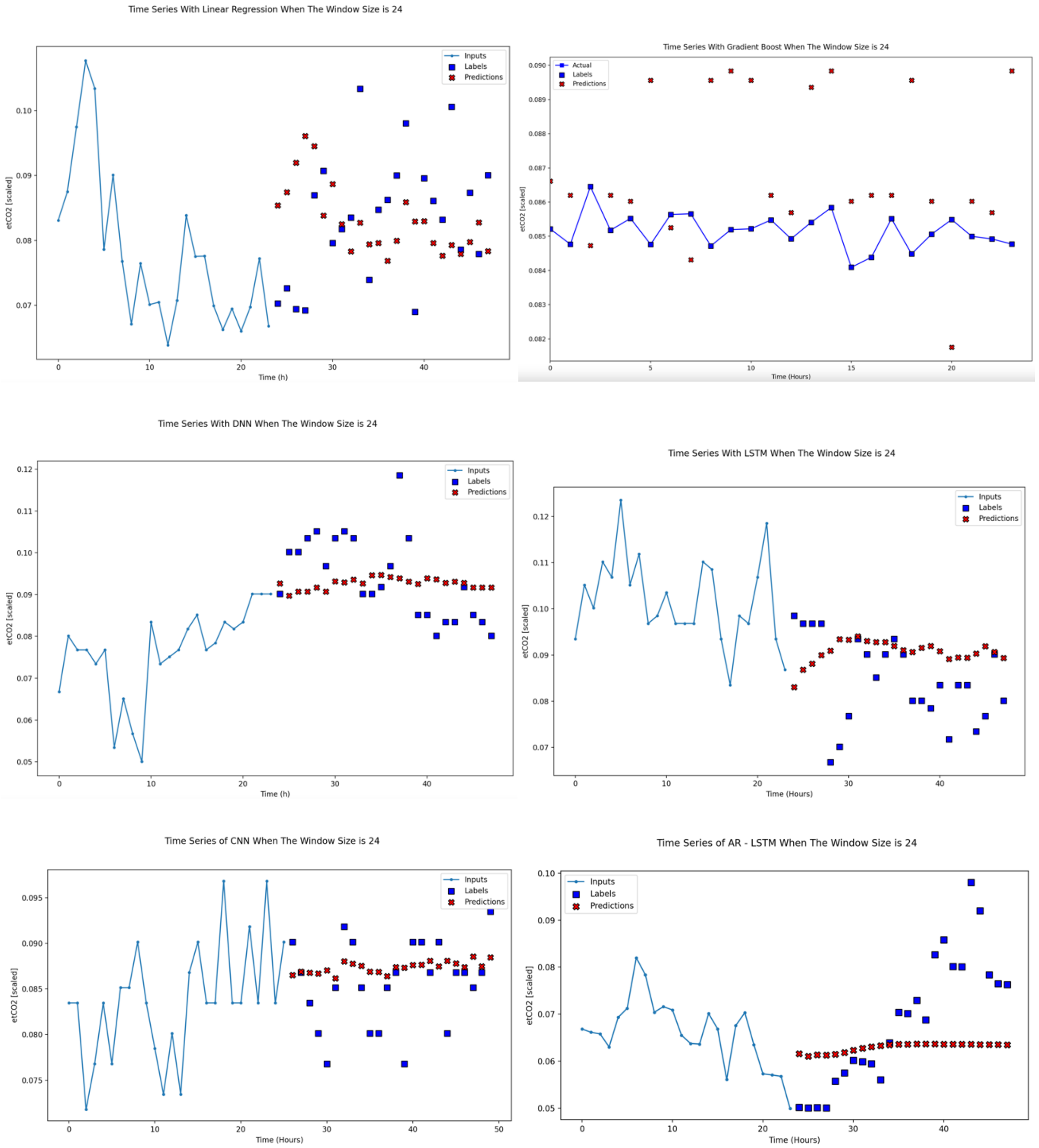


*Figure 25: Histogram of the MAPE  of all the models with window size 24, feature noise percentage of 2% and five selected features: Oxgyen_etCO2_ratio, FiO2, DOL, Age_of_diagnois and BW.*

Lower MAPE values reflect stronger model performance, indicating that the predictions are more aligned with actual observations. For the baseline model, the MAPE depicted in Figure 25 was 26.2% for the validation set and 17.6% for the test set—these were the highest values recorded among all models, which is expected since the baseline model usually offers simplicity and interpretability rather than precision. The XGBoost model outperformed all others, recording MAPEs of 4.8% for validation and 5.37% for testing, making it roughly five times more accurate than the baseline and surpassing all other models evaluated in this study.

The CNN and DNN models demonstrated the second-best accuracy with MAPEs of 17.6% for validation and 12.1% for testing, showcasing a performance that was 1.5 times better to the baseline. On the other end of the spectrum, the LSTM and AR-LSTM models exhibited the least favourable performance, with accuracy closely mirroring that of the baseline. It can be confidently seen that none of the models in the figure displayed any signs of overfitting, as the validation MAPE was relatively close to the test MAPE for all the models in this study.

In a medical context, to understand if these MAPE values are sufficiently low for clinical application, research into machine learning's efficacy had to be evaluated. The particular study – “Machine learning’s efficiency in disease diagnosis and prognosis within plastic surgery - indicated that high accuracy levels are between 88.80% and 80.28%—which translates to MAPEs of 11.2% to 19.7%. Given that the best-performing model, XGBoost, achieved a MAPE of 4.89%, it is reasonable to suggest that this model possesses a level of accuracy that is suitable for certain medical applications. While no specific MAPE benchmark for neonatal care was identified, it is cautiously assumed that the observed accuracy is satisfactory for medical usage in that context as well. Consequently, XGBoost was selected as the model for further refinement and development into an explainable machine-learning model for potential clinical deployment. It’s essential to keep in mind that the "best" performing model is highly dependent on the dataset and features. The best model may not provide equally satisfying performance for another dataset related to the same research problem.



*Figures 26, 27, 28, 29, 30, 31: The Scatter plot of the inputs, labels and predictions of etCO2 for the Linear Regression, XGBoost, DNN, LSTM, CNN and AR-LSTM respectively, for a time window of 24 hours.*

The graphs presented in Figures 26-31 offer a visual comparison of the predicted and actual scaled etCO2 values for all models. These graphs can help identify patterns in the data. Figure 26 shows the results for the linear model, which displays a significant difference between the predicted and actual values, indicating low predictive power. In contrast, Figure 27, which represents the XGBoost model, shows a tight grouping of predictions around the actual labels, suggesting high accuracy and the lowest MAPE among the models compared. Figure 28, the DNN model, demonstrates moderate alignment between the predictions and actual labels, but there are visible discrepancies, particularly for higher etCO2 values. The LSTM model in Figure 29 shows reasonable predictive accuracy with some outliers, particularly for lower etCO2 values. Similarly, CNN in Figure 30 performs similarly to LSTM, with slightly better predictions at lower etCO2 values. Lastly, the AR-LSTM model in Figure 31 displays consistent predictions but struggles to capture the peaks in the etCO2 values.

**5.6 Hyper-tunning parameter for the selected model:**

Hyperparameter tuning is a crucial step in the model development process, aimed at enhancing predictive accuracy by optimising the model's parameters. The XGBoost model was selected due to its lower predictability of 4.8% for the validation set. The primary parameters subject to tuning include ‘max_depth’, controlling the depth of the trees and thus the model's complexity; ‘learning_rate’, which affects the speed and performance of the model; ‘n_estimators’, which is the number of trees in the ensemble; and ‘colsample_bytree’, which determines the fraction of features considered for each tree. To hyper tune parameters, the algorithm selected to find the optimal parameters was Grid Search, an exhaustive search. It involves a cross-validated grid search over a parameter grid [34], exploring a combination of parameters to find the most effective combination. The best combination is then evaluated through its accuracy score MAPE. This approach is expected to refine the model's performance on unseen data, ensuring the model's robustness and reliability in a medical context.

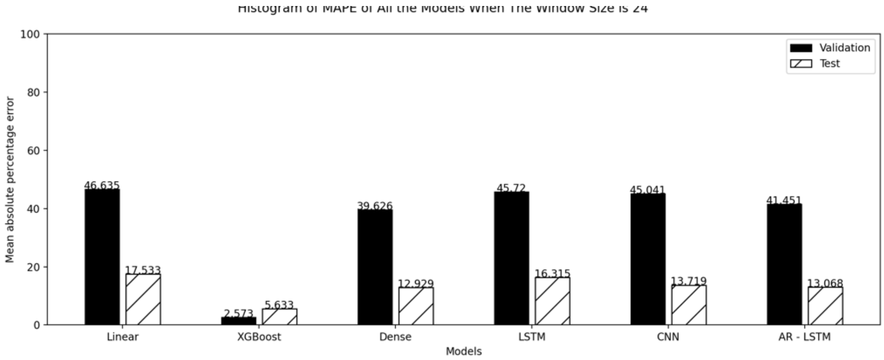


*Figure 31: Histogram of MAPE of all the models after GridSearch on the XGboost Model*

After experimenting with different parameter ranges on the grid search for 'max_depth', 'learning_rate', 'n_estimators', and 'colsample_bytree', it was found that the best parameters were ‘colsample_bytree’ at 0.5, ‘learning_rate’ at 0.05, ‘max_depth’ at 3, and ‘n_estimators’ at 1500. Figure 31 shows that by using grid search to find the best parameters, the MAPE for the validation set decreased from 4.8% to 2.6%. However, the MAPE validation remained the same at 5.6% regardless of the parameter range tweaks for the grid search.

5.7 XAI using SHAP

The objective of this study was to develop an explainable machine-learning model that is capable of predicting CO2 levels to enhance respiratory support. So far, an accurate and high-performing model has been constructed using the XGBoost algorithm. However, XGBoost models can often be seen as "black boxes" due to their intricate structure which makes predictions difficult to explain transparently. To understand the inner workings of this model and assess the influence of individual features on the prediction of etCO2, a SHAP library in Python was used. This approach allows for an analysis of feature impacts, specifying which features are most significant in predicting end-tidal CO2 and the extent of each feature’s influence.

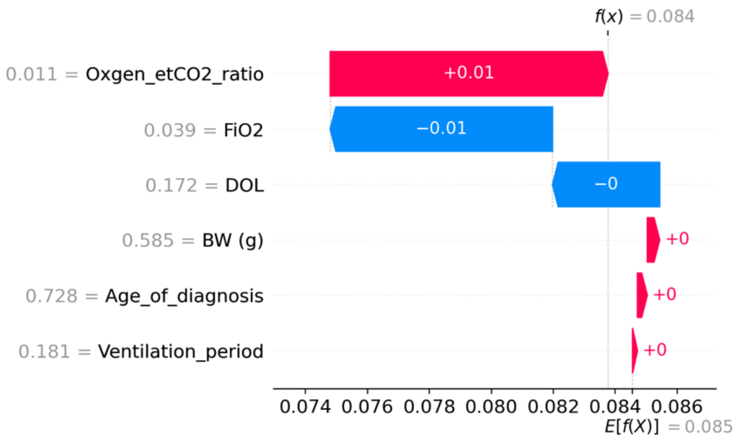


*Figure 32: SHAP Value Impact on Model Prediction*

Figure 32 is a SHAP value plot that interprets the influence of each feature on the model's prediction of CO2 levels in neonates. The plot provides a visual representation of the size and direction of each feature's effect on the predicted outcome relative to the base expectation (E[f(X)] = 0.085).

From this plot, it is observed that 'Oxygen_etCO2_ratio' has a positive effect on the model's prediction by increasing it by 0.01, suggesting that as the ratio increases, so does the predicted CO2 level. Conversely, 'FiO2' has a negative effect by the same magnitude, indicating that higher FiO2 values are associated with a decrease in the predicted CO2 level.

It's important to note that the changes shown here are moderate in the context of the data's scale and precision, which is typically accurate to two decimal places. Features such as 'BW (g)', 'Age_of_diagnosis', and 'Ventilation_period' show very small SHAP values, almost negligible, denoting a minimal individual impact on the predicted value.

The right side of the plot features a vertical line marked with f(x) = 0.084, which is the actual predicted outcome from the model for the specific case being analysed. The divergence of this value from the base value E[f(X)] indicates the cumulative impact of the features on the prediction.

# SECTION 5:

This section focuses on professionalism and aims to reflect on the wider professional responsibilities, codes, and regulations associated with the study. The report discusses the social, economic, legal, and ethical impacts of the study in the context of the medical field. It assesses the benefits of this project and identifies if it can be further improved for future use in medical care.

# 6 PROFESSIONALISM AND RESPONSIBILITY:

## 6.1 SOCIAL, ECONOMIC, LEGAL IMPACTS:

This project not only aspires to fulfil healthcare workers' needs, as guided by the customer orientation principle in ISO 90003, but also contributes to the health and well-being of neonates in accordance with social responsibility principles outlined in ISO 26000. However, due to the novelty of machine learning and XAI in healthcare, there are numerous concerns regarding their accuracy in implementation in this field, as studies show that doctors struggled to trust AI for several reasons [35]. As research progresses and the development of these technologies improves, XAI will be seamlessly integrated into healthcare practices and will be more widely accepted and trusted in medical practices.

Currently, the model trained and the XAI created may serve as a cost-effective minimum viable product (MVP). With additional data, expanded domain expertise, and further development, there's significant potential for this MVP to evolve into a comprehensive product service, offering substantial economic value to the healthcare sector. Currently, there is very little legal implementation of machine learning and XAI (Naik et al., 2022). As XAI is used in more healthcare systems, it is expected that various changes and the possibility of regulatory intervention will occur. It is possible that governmental bodies may introduce new rules and regulations to govern the use of XAI in medical settings.

## 6.2 TECHNOLOGICAL TRENDS AND ENVIRONMENTAL ISSUES:

The approach taken for the software development of the models in this study reflects the principles within ISO 90003, as the models were created systematically and prioritised quality-driven results. To fully comply with the standards in ISO 90003 guidelines, a quality-documented process was implemented, and records were maintained. The software developed in this study underwent continuous validation and verification to ensure it remains safe, effective, and reliable. Additionally, this study leveraged widely accessible technology trends, such as libraries on machine learning model tools develop the models in this study.

Regarding the environmental impacts, this project is expected to have a minimal effect. Although there may be higher computing demands as the dataset grows and complexity increases, it is anticipated that the environmental footprint will remain relatively low. It's worth noting that the data collection took place in London, but the project is not restricted to any particular area and can be adjusted to various regions and environments.

## 6.3 BENEFITS AND DRAWBACKS:

The project has the potential to have significant positive impacts, such as reducing the amount of time neonates require mechanical respiration and mitigating the risks of hypocarbia and hypercarbia. The use of XAI in medicine is expected to enhance clinical decision-making, allowing for more personalised and accurate patient care that can lead to shorter ICU stays, and lower risks of developing chronic lung disease. The clear, understandable insights provided by XAI allow healthcare professionals to make informed decisions, potentially leading to better healthcare outcomes for neonates.

However, using XAI has negative impacts, such as inaccurate predictions [35] that could result in inadequate care or even increased mortality rates among neonates. Moreover, the use of XAI within clinics can lead to complacency in monitoring and responding to neonatal conditions due to human nature, which is a concern. Additionally, the medical field remains cautious about XAI, so there is an emphasis on the need for a careful and informed approach to its integration into healthcare practices.

### 6.4 ETHICS AND CODE OF CONDUCT:

As the use of artificial intelligence (AI) becomes more common in healthcare, particularly in neonatal care, it is essential to consider the ethical implications. The use of AI in neonatal care raises concerns about patient privacy, data security, and algorithmic bias. Electronic medical data used in AI-based systems can lead to ethical issues such as privacy breaches, unfairness, lack of accountability, and transparency[19]. Before handling and manipulating any data in this study ethical approval was obtained from Prof. Theodore Dassios, who works in the Department of Women & Children's Health at King's College London. The patient data that was used in this project was anonymous to protect privacy, which is a direct accordance with the commitment to protect the public interest and data protection as per the BCS guidelines and 'The EU 's General Data Protection Regulation (GDPR)" which allows individuals the right to control over their personal data (Ethics Guidelines for Trustworthy AI | Shaping Europe's Digital Future, n.d.).To ensure fairness, the dataset was thoroughly examined for any data that may lead to bias or discrimination based on factors such as nationality, ethnicity and geographical location. After a thorough examination, it was determined that the data does not exhibit any observable signs of bias. Accountability was implemented by clearly defining responsibility for the model's outcomes. However, this was difficult because with machine learning, the algorithm is not capable of predicting the future and cannot be held morally responsible for any complications that may result due to its use[36]. Transparency is very important for Machine learning developers as they have an ethical responsibility to record and report the model's performance metrics appropriately. In this study, transparency was exercised by displaying all the graphs, tables and code used to implement the models. However, even with this transparency, interpretations of Machine learning models can pose an ethical challenge to developers when the model is a "black box" due to its high complexity, such as XGBoost.

# 6 CONCLUSION

This project investigated the application of machine learning and explainable AI (XAI) in monitoring CO2 levels in neonates requiring respiratory support. The objective was to develop a predictive tool that not only forecasts CO2 levels with high accuracy but also elucidates these predictions for medical practitioners in an understandable manner.

The research commenced by identifying the challenges faced in current neonatal respiratory support practices, particularly the difficulty in accurately monitoring CO2 levels. Employing various methodological approaches, including time series forecasting and machine learning models, with a focus on the XGBoost algorithm, the study achieved significant predictive accuracy. The performance of this model was further enhanced through feature engineering, data augmentation, and hyperparameter tuning.

A key achievement of this study was the incorporation of SHAP values to interpret the decision-making process of the predictive model. This approach enabled the identification of key features influencing CO2 level predictions, thereby fostering trust among medical practitioners and encouraging the integration of this tool into clinical settings.

The findings suggest that machine learning and XAI hold considerable promise for improving the efficiency and effectiveness of neonatal respiratory support. The developed predictive tool has the potential to streamline the management of respiratory support, possibly reducing the need for prolonged mechanical ventilation and the associated risks.

Future directions for this research include expanding the dataset to improve the model's accuracy and scalability. Collaboration with domain experts is recommended to refine the model's applicability in clinical settings, ensuring that its benefits extend from theoretical to practical applications in neonatal care.

In summary, this study represents a significant advancement towards integrating artificial intelligence into medicine, specifically in neonatal care. By leveraging machine learning and explainable AI, there is potential for more personalized, accurate, and effective care for newborns requiring respiratory support.

### 6.4 POSSIBLE IMPROVEMENTS:

To improve the current model in this study and possibly implement it in a clinical setting in the future, certain improvements must be made, such as expanding the dataset for improved scalability and accuracy and integrating expertise from domain specialists and research professionals. The potential risk of misusing the outcomes from this study's limited datasets without further validation and research could be very grave in medical care. Therefore, further engagement with medical professionals who bring more domain knowledge is necessary, ensuring that the project's findings are not only scientifically sound but also practically applicable and beneficial in the real-world context of neonatal care.

## 7 REFERENCES:

[1] I. M. Cheifetz, "Invasive and Noninvasive Pediatric Mechanical Ventilation," *Respir Care*, vol. 48, no. 4, 2003.

[2] R. P. Anne and S. Murki, "Noninvasive Respiratory Support in Neonates: A Review of Current Evidence and Practices," *Indian J Pediatr*, vol. 88, no. 7, p. 670, Jul. 2021, doi: 10.1007/S12098-021-03755-Z.

[3] "Overview of initiating invasive mechanical ventilation in adults in the intensive care unit - UpToDate." Accessed: Apr. 10, 2024. [Online]. Available: https://www.uptodate.com/contents/overview-of-initiating-invasive-mechanical-ventilation-in-adults-in-the-intensive-care-unit/print

[4] G. Rocha *et al.*, "Respiratory Care for the Ventilated Neonate," *Can Respir J*, vol. 2018, 2018, doi: 10.1155/2018/7472964.

[5] R. Pauldine, G. Beck, J. Salinas, and D. W. Kaczka, "Closed-Loop Strategies for Patient Care Systems The Journal of TRAUMA Injury, Infection, and Critical Care," 2008, doi: 10.1097/TA.0b013e31816bce43.

[6] D. Sankaran, L. Zeinali, S. Iqbal, P. Chandrasekharan, and S. Lakshminrusimha, "Non-invasive carbon dioxide monitoring in neonates: methods, benefits, and pitfalls," *Journal of Perinatology*, vol. 41, no. 11, p. 2580, Nov. 2021, doi: 10.1038/S41372-021-01134-2.

[7] O. Hochwald, L. Borenstein-Levin, G. DInur, H. Jubran, S. Ben-David, and A. Kugelman, "Continuous Noninvasive Carbon Dioxide Monitoring in Neonates: From Theory to Standard of Care," *Pediatrics*, vol. 144, no. 1, 2019, doi: 10.1542/PEDS.2018-3640.

[8] G. A. Hawkes *et al.*, “Delivery room end tidal CO 2 monitoring in preterm infants <32 weeks”, doi: 10.1136/archdischild-2015-308315.

[9] S. K. Wong *et al.*, “Carbon dioxide levels in neonates: what are safe parameters?,” *Pediatr Res*, vol. 91, no. 5, p. 1049, Apr. 2022, doi: 10.1038/S41390-021-01473-Y.

[10] “End-tidal carbon dioxide,” 2008.

[11] K. B. Johnson *et al.*, “Precision Medicine, AI, and the Future of Personalized Health Care,” *Clin Transl Sci*, vol. 14, no. 1, pp. 86–93, Jan. 2021, doi: 10.1111/CTS.12884.

[12] J. M. Arnal *et al.*, “Feasibility study on full closed-loop control ventilation (IntelliVent-ASV$^{TM}$) in ICU patients with acute respiratory failure: A prospective observational comparative study,” *Crit Care*, vol. 17, no. 5, pp. 1–10, Sep. 2013, doi: 10.1186/CC12890/TABLES/5.

[13] A. A. Chakkarapani *et al.*, “‘Current concepts of mechanical ventilation in neonates’ – Part 1: Basics,” *Int J Pediatr Adolesc Med*, vol. 7, no. 1, p. 13, Mar. 2020, doi: 10.1016/J.IJPAM.2020.03.003.

[14] G. Schmalisch, “Current methodological and technical limitations of time and volumetric capnography in newborns,” *Biomed Eng Online*, vol. 15, no. 1, pp. 1–13, Aug. 2016, doi: 10.1186/S12938-016-0228-4/FIGURES/5.

[15] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G. Z. Yang, “XAI-Explainable artificial intelligence,” *Sci Robot*, vol. 4, no. 37, Dec. 2019, doi: 10.1126/SCIROBOTICS.AAY7120/ASSET/635AF7F2-CA10-40DD-87D2-7DCD424AE2CD/ASSETS/GRAPHIC/AAY7120-F1.JPEG.

[16] T. C. Kwok *et al.*, “Application and potential of artificial intelligence in neonatal medicine,” *Semin Fetal Neonatal Med*, vol. 27, no. 5, Oct. 2022, doi: 10.1016/J.SINY.2022.101346.

[17] J. F. Hsu *et al.*, “Machine learning algorithms to predict mortality of neonates on mechanical intubation for respiratory failure,” *Biomedicines*, vol. 9, no. 10, Oct. 2021, doi: 10.3390/BIOMEDICINES9101377/S1.

[18] E. Keles and U. Bagci, “The past, current, and future of neonatal intensive care units with artificial intelligence: a systematic review,” *NPJ Digit Med*, vol. 6, no. 1, Dec. 2023, doi: 10.1038/S41746-023-00941-5.

[19] T. Basu, S. Engel-Wolf, and O. Menzer, “The Ethics of Machine Learning in Medical Sciences: Where Do We Stand Today?,” *Indian J Dermatol*, vol. 65, no. 5, p. 358, Sep. 2020, doi: 10.4103/IJD.IJD_419_20.

[20] M. Xie, N. Jean, M. Burke, D. Lobell, and S. Ermon, “Transfer learning from deep features for remote sensing and poverty mapping,” *30th AAAI Conference on Artificial Intelligence, AAAI 2016*, pp. 3929–3935, 2016, doi: 10.1609/aaai.v30i1.9906.

[21] Z. Zhu, T. Liu, G. Li, T. Li, and Y. Inoue, “Wearable Sensor Systems for Infants,” *Sensors 2015, Vol. 15, Pages 3721-3749*, vol. 15, no. 2, pp. 3721–3749, Feb. 2015, doi: 10.3390/S150203721.

[22] M. Peixeiro, “Time Series Forecasting in Python.”

[23] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System”, doi: 10.1145/2939672.2939785.

[24] T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, pp. 785–794, Mar. 2016, doi: 10.1145/2939672.2939785.

[25] E. A. Houssainy Rady, H. Fawzy, and A. Mohamed Abdel Fattah, “Time Series Forecasting Using Tree Based Methods,” *Journal of Statistics Applications & Probability An International Journal*, vol. 10, no. 1, p. 229, 2021, doi: 10.18576/jsap/100121.

[26] M. Ali, A. Dewan, A. K. Sahu, and M. M. Taye, “Understanding of Machine Learning with Deep Learning: Architectures, Workflow, Applications and Future Directions,” *Computers 2023, Vol. 12, Page 91*, vol. 12, no. 5, p. 91, Apr. 2023, doi: 10.3390/COMPUTERS12050091.

[27] S. Hochreiter, “Recurrent Neural Net Learning and Vanishing Gradient,” *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 6, no. 2, pp. 107–116, 1998, Accessed: Apr. 11, 2024. [Online]. Available: http://www7.informatik.tu-muenchen.de/~hochreit

[28] M. Norman, B. Jonsson, J. Söderling, L. J. Björklund, and S. Håkansson, “Patterns of Respiratory Support by Gestational Age in Very Preterm Infants,” *Neonatology*, vol. 120, no. 1, pp. 142–152, Mar. 2023, doi: 10.1159/000527641.

[29] R. L. Dellaca, C. Veneroni, and R. Farre, “Educational aims,” vol. 13, no. 2, 2017, doi: 10.1183/20734735.007817.

[30] Z. Khazaei *et al.*, “Risk Factors Associated with Low Birth Weight Among Infants: A Nested Case-Control Study in Southeastern Iran,” *Int J Prev Med*, vol. 12, no. 1, 2021, doi: 10.4103/IJPVM.IJPVM_300_20.

[31] D. Muktan, R. R. Singh, N. K. Bhatta, and D. Shah, “Neonatal mortality risk assessment using SNAPPE-II score in a neonatal intensive care unit,” *BMC Pediatr*, vol. 19, no. 1, pp. 1–4, Aug. 2019, doi: 10.1186/S12887-019-1660-Y/FIGURES/1.

[32] G. S. Handelman *et al.*, “Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods,” *American Journal of Roentgenology*, vol. 212, no. 1. American Roentgen Ray Society, pp. 38–43, Jan. 01, 2019. doi: 10.2214/AJR.18.20224.

[33] “3.1. Cross-validation: evaluating estimator performance — scikit-learn 1.4.2 documentation.” Accessed: Apr. 11, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/cross_validation.html

[34] “3.2. Tuning the hyper-parameters of an estimator — scikit-learn 1.4.2 documentation.” Accessed: Apr. 11, 2024. [Online]. Available: https://scikit-learn.org/stable/modules/grid_search.html#exhaustive-grid-search

[35] M. I. Ahmed, B. Spooner, J. Isherwood, M. Lane, E. Orrock, and A. Dennison, “A Systematic Review of the Barriers to the Implementation of Artificial Intelligence in Healthcare,” *Cureus*, vol. 15, no. 10, Oct. 2023, doi: 10.7759/CUREUS.46454.

[36] N. Naik *et al.*, “Legal and Ethical Consideration in Artificial Intelligence in Healthcare: Who Takes Responsibility?,” *Front Surg*, vol. 9, p. 862322, Mar. 2022, doi: 10.3389/FSURG.2022.862322/BIBTEX.

[37] “Ethics guidelines for trustworthy AI | Shaping Europe’s digital future.” Accessed: Apr. 09, 2024. [Online]. Available: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

# 9 APPENDIX:

### 2.7 Challenges in Developing Closed-Loop $CO_2$ Systems

Developing closed-loop CO2 systems for neonatal care is a complex task that involves various technical and clinical considerations. Implementing closed-loop CO2 systems in neonatal care involves overcoming significant challenges to ensure they are safe, effective, and clinically viable. Precision in CO2 monitoring is paramount, with TcCO2 technology leading the way, yet guaranteeing sensor accuracy for neonates remains difficult.

Clinical integrating closed-loop CO2 systems into clinical practice requires collaboration between engineers and clinicians. Clinical trials are necessary to evaluate the safety and effectiveness of closed-loop CO2 systems in the real world. These trials should focus on patient outcomes, and should be accepted by healthcare practitioners.

Moreover, the closed-loop control algorithms must be meticulously designed and tested to maintain CO2 levels within a narrow therapeutic range (Katheria et al., 2019).