Received 14 June 2022; revised 10 August 2022 and 23 September 2022; accepted 21 October 2022.
Date of publication 26 October 2022; date of current version 4 November 2022.

Digital Object Identifier 10.1109/JTEHM.2022.3217428

Res-SE-ConvNet: A Deep Neural Network for Hypoxemia Severity Prediction for Hospital In-Patients Using Photoplethysmograph Signal

TALHA IBN MAHMUD, (Graduate Student Member, IEEE), SHEIKH ASIF IMRAN, AND CELIA SHAHNAZ, (Senior Member, IEEE)

Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology (BUET), Dhaka 1205, Bangladesh

CORRESPONDING AUTHOR: T. I. MAHMUD (talhaibnmahmud30@gmail.com)

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by the Institutional Review Boards of BIDMC and the Massachusetts Institute of Technology.

ABSTRACT Determining the severity level of hypoxemia, the scarcity of saturated oxygen (SpO2) in the human body, is very important for the patients, a matter which has become even more significant during the outbreak of Covid-19 variants. Although the widespread usage of Pulse Oximeter has helped the doctors aware of the current level of SpO2 and thereby determine the hypoxemia severity of a particular patient, the high sensitivity of the device can lead to the desensitization of the care-givers, resulting in slower response to actual hypoxemia event. There has been research conducted for the detection of severity level using various parameters and bio-signals and feeding them in a machine learning algorithm. However, in this paper, we have proposed a new residual-squeeze-excitation-attention based convolutional network (Res-SE-ConvNet) using only Photoplethysmography (PPG) signal for the comfortability of the patient. Unlike the other methods, the proposed method has outperformed the standard state-of-art methods as the result shows 96.5% accuracy in determining 3 class severity problems with 0.79 Cohen Kappa score. This method has the potential to aid the patients in receiving the benefit of an automatic and faster clinical decision support system, thus handling the severity of hypoxemia.

INDEX TERMS Saturated oxygen, attention, feature map, excitation, deep learning.

I. INTRODUCTION

Oxygen saturation (SpO2) blood is measured by the ratio between the concentration of hemoglobin which have formed a chemical compound with oxygen, called oxy-hemoglobin, and the total concentration of hemoglobin. In human body, standard values of oxygen saturation are above 96% [1].

Hypoxemia is the state when the saturated oxygen level of patient falls generally below 90% [2], a condition which might be symptom of diseases like asthma or lungs tumor [3]. It can be a dangerous issue and patients of high risk are often transferred immediately to the Intensive Care Unit (ICU) for close monitoring and rapid intervention [4]. Hypoxemia is a common sedation-related complication [5]. Although it normally remains in mild state, and spontaneous recovery is likely, hypoxemia remains the principal cause of increased morbidity and mortality [6], which in turn may become lethal and require immediate medical attention. It is even the most common complication of tracheal intubation in ICU [7], [8], [9], [10] and is associated with cardiac arrest [7], [11], [12]. Avoidance of hypoxemia during tracheal intubation is a goal in clinical practice [13]. Therefore, early warning and a reliable method of risk stratification for hypoxemia may help the physician select patients who would benefit most from an aggressive intervention and thereby confirm the optimum utilization of the medical resource allocation [4], [14].

The detection of hypoxemia is highly dependant on the detection of current state of saturated oxygen level of the patient which is widely measured by pulse oximeter using dual wavelength Photoplethysmography (PPG) [15]. Takuo Aoyagi is the pioneer to design pulse oximetry in 1971 by using the ratio of red to infrared light absorption of pulsating components at the measuring area [16]. The standard of care for the administration of a general anesthetic in the U.S. included pulse oximetry and the application of the device spread from the operating room to recovery rooms, and then to ICUs. It was of particular value in the neonatal unit.

However despite the wide application of this device in the hospital as well as the household, its high sensitivity may lead to high rate of false alarms [17]. As a result it can desensitize the care givers to real emergencies [18], [19]. Therefore, an alternate approach should be pursued to compensate the sensitivity.

Different machine learning techniques such as support vector machine and artificial neural network were applied to predict SpO2 using blood visible spectra during ex-vivo treatments [1]. A prediction model of Hypoxemia was designed by Geng et al. [14] using demographic data, concurrent chronic disease information, anesthetic dose and Modified Observer's Assessment of Alertness/Sedation (MOAA/S) scores. McKown et al. [13] developed logistic regression model to predict severe hypoxemia. An artificial neural network model was designed in [20] where body mass index, neck circumference and data of habitual snoring were used as input to predict hypoxemia. Although these novel approaches are promising, they require several patient data for prediction. Additionally, to facilitate medical resources, it is extremely crucial to classify the hypoxemia patients in terms of their severity level which is especially important during the outbreak of covid variants. In [21], Ghazal et al. used machine learning approaches such as artificial neural network (ANN) and bootstrap aggregation of complex decision trees (BACDT) to evaluate the severity level of the patient. But their method had to use several patient data in addition to continuous biomedical signals to predict the outcome.

In this paper, we propose a new residual Squeeze and Excitation (SE) Attention based convolutional neural network that can predict the severity level of hypoxemia of a critical patient using only PPG signal. Rather than feeding the signal only into a stack of convolutional layers, a residual approach of SE attention based parallel branch is proposed where the extracted features can be imposed on the traditional convolutional output to generate more fine-tuned parameters. The result of the model is further compared with conventional machine learning classification approaches along with the existing deep neural architectures. The proposed model has the potential to aid the physicians in rapid classification of the patients on the basis of their need of intensive care in time of urgency.

II. PROPOSED METHODOLOGY

The proposed methodology is divided into several section. At first the pre-processing of the extracted input data is explained. Then, the necessity and procedure of data sampling is described. Later on, the novel neural architecture, called "Res-SE-ConvNet" with function of its individual blocks is demonstrated with necessary flow charts. Finally, the loss function necessary for model optimization is explained with proper detail.

A. DATA PRE-PROCESSING

The digitized PPG data collected from the patients with corresponding SpO2 value are at first divided into several frames with fixed frame length to facilitate the processing of the network. After that, the constructed frames are annotated into 3 separated labels depending on their oxygen saturation value for evaluation purposes. A patient having SpO2 level of greater than 91% may not need immediate medical attention whereas patients with SpO2 level between 91% to 85% should be provided with necessary medical attention. If the oxygen level drops below 85%, then the case should be considered as critical, and the patient needs immediate medical procedure to be resuscitated to normal condition. To this goal, the frames are labelled as 0, 1 and 2 accordingly depending on the oxygen label- 0 being normal (greater than 91%), 1 being moderate (85% - 91%) and 2 being critical (less than 85%) [21].

Let us consider the whole set of extracted PPG frame set to be denoted as

$$D = \{(\mathbf{x}_i, y_i) |\; i = 0, 1, 2, \ldots, N-1\} \tag{1}$$

whereas N is the total number of frames, $\mathbf{x}_i$ is the i_{th} frame of predefined length and $\mathbf{y}_i$ is its corresponding annotated label. All the frames are extracted from the raw PPG signal $\mathbf{X}$ and its corresponding annotation vector $\mathbf{Y}$:

$$x_i = X[(1 + s * i), (2 + s * i), \ldots, (l + s * i)] \tag{2}$$

$$y_i = Y[i] \qquad \forall i \in \{0, 1, 2, 3, \ldots, N-1\} \tag{3}$$

where, s is the frame shift of the raw frame. As we did not want any data overlapping between two frames, the frame shift was set to be equal to frame length. The division of continuous PPG signal to generate pre-processed PPG frames of fixed length of 1 second can be viewed in Figure 2. For case of simplicity only a 30 second of duration of the whole PPG signal has been chosen to demonstrate its division into 3 hypoxemia classes according to their SpO2 level.

B. DATA SAMPLING

After the frame creation, it can be observed that the data contain a ratio of P:Q:R among the normal, moderate and severe classes, whereas $P \gg Q \approx R$, a heavy imbalance due to the extreme scarcity of moderate and critical frames. Training these data frames directly to any network will have the tendency to be overfit on the normal class. Due to the huge number of frames, the model accuracy might be quite high, but these performance can not be acceptable in realistic point of view. To train the model to detect all kinds of labels, the dataset must be balanced for all these 3 classes. For this purpose, a combination of up-sampling and down-sampling of relevant classes was necessary before designing the neural network. Frames of moderate and critical classes were fed to the Adaptive Synthetic (ADASYN) technique [22] to adaptively generate minority data frame while paying attention to their density distribution. At the same time, frames of normal class were fed to random under-sampler operation instead of Tomek Links to avoid the risk of discarding potential data as borderline samples can be important in specifying the decision border [23]. By following the operation, a balanced database was generated for robust model performance. After the operation the ratio became 1:1:1 for all the classes and these newly sampled frames were used to train the neural network. After performing these sequential two pre-processing,

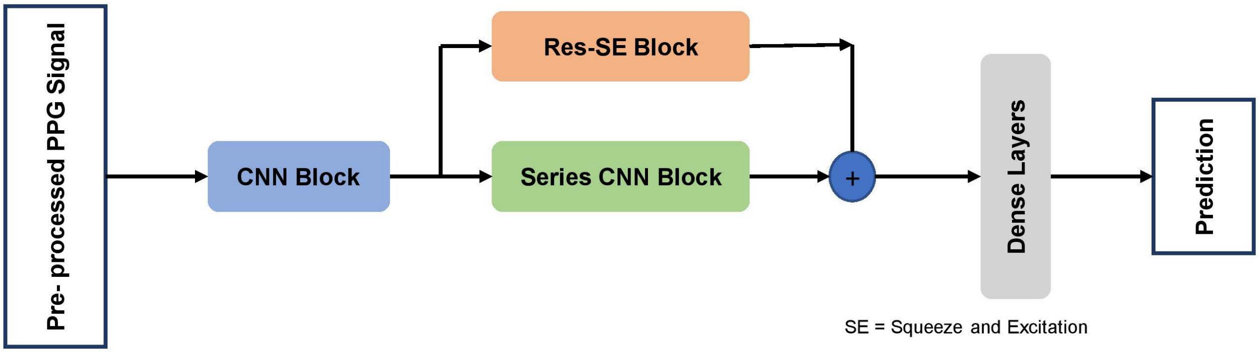

FIGURE 1. Schematic representation of the proposed workflow.

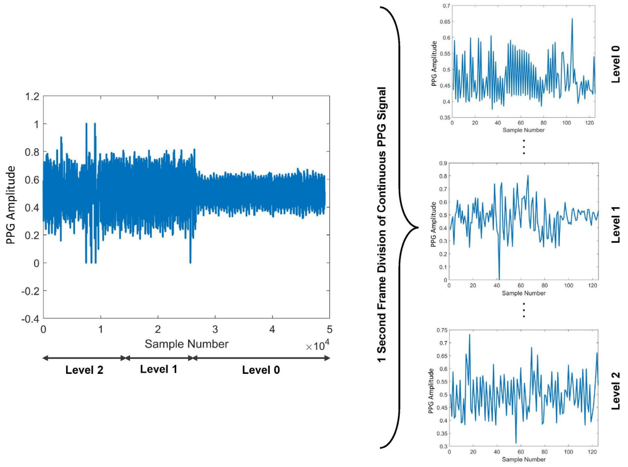

FIGURE 2. Pre-processing of PPG signal to generate fixed 1 second length of frames.

the generated PPG frames were ready to be used to train and evaluate the network.

C. PROPOSED DEEP NEURAL ARCHITECTURE

As shown in Fig. 1, the proposed methodology is divided into four main sections namely- Convolutional Neural Network (CNN) block, Series CNN Block, Rsidual Squeeze and Excitation Attention (Res-SE) Block and Dense Layers. Firstly, the pre-processed signal is fed into the CNN block for feature extraction. The output feature map is then used as input to two individual sub-networks. The Series CNN Block continues to fine tune the features while the Residual Squeeze and Excitation Attention (Res-SE) Block quantifies the interdependence of each node of the feature map to the output. The feedback of these two routes is then merged and converted to a flattened feature vector to be processed with a series of densely connected layers to converge towards the final prediction of hypoxemia label. Detailed architectural analysis of each sub-network is provided in the following discussion.

1) CONVOLUTIONAL NEURAL NETWORK (CNN) BLOCK

The name of the block is called as CNN block because of the primary CNN layer that resides in the segment although the CNN layer is not the only layer this block contains. The output of CNN layer is fed into a one dimensional maxpooling layer along with PRelu activation function prior to batch normalization. The description of the whole block is explained here:

Each CNN block contains 1 trainable convolutional layer with kernel size of 3 and channel number of 64. The 2D feature map is sub-sampled to reduce the number of parameters to be computed using MaxPool layer with pool size of 2. In all the SCNN blocks, Parametric Rectified Linear Unit (PRelu) is used as non-linear activation function for faster convergence where PRelu is:

$$\mathrm{PRelu(x)} = \begin{cases} x, & \text{if } x \geq 0 \\ \alpha x, & \text{otherwise} \end{cases} \tag{4}$$

Here, α is the slope for mapping the negative value of the input whose value, for our proposed method, was chosen to be 0.2. As the value of the negative slope is made constant for the whole model training, the PReLU acts as Leaky ReLU and thereby eradicates the dead neuron problem [24] that ReLU activation can create by turning the neurons to off state if they are not activated initially.

Lastly, a batch normalization process is performed on the feature map to avoid overfitting of the model. The complete block can be viewed in Fig. 3 where the dimension of x_{in} can be altered randomly and the output feature map x_{out} will have the same length as the input but the channel number will be 64. For our method the input length was selected to be 125.

2) SERIES CNN BLOCK

This block is made up of a series of CNN block to allow a hierarchical decomposition of the input data which can be seen at Fig. 3b. Each stack of repetitive CNN layers helps the network extract relevant information from spatial local feature map, resulting in creating deeper representation of the input than the previous one and thereby improve the performance of the model at a low computational cost as a whole. The number of CNN block in this route can be varied to analyze the performance of the proposed model. Smaller number of blocks typically denotes lesser number of suitable features that will result in performance degradation. Higher number of blocks might be able to achieve better result at

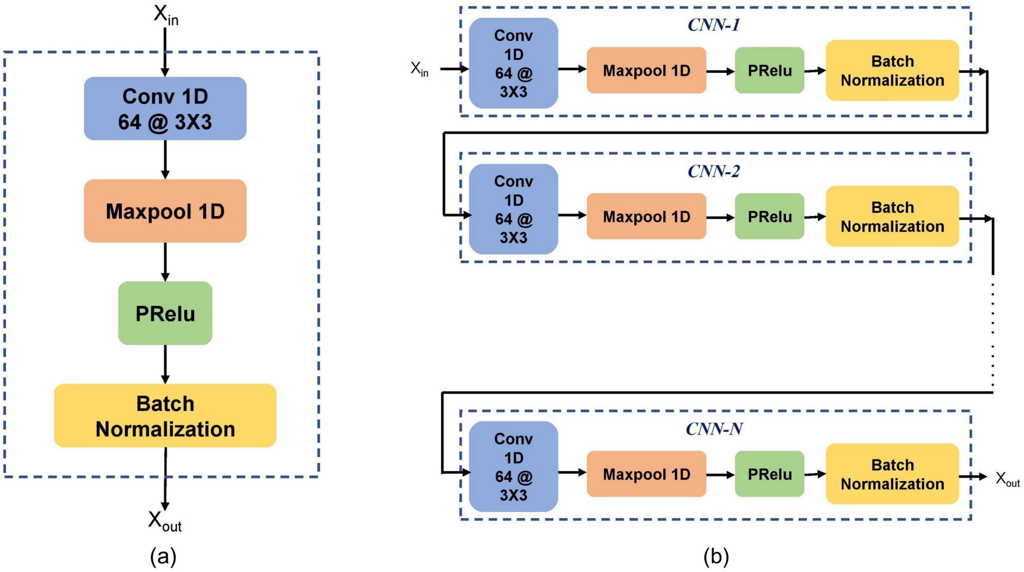

FIGURE 3. Schematic representation of (a) CNN block, and (b) Series CNN block.

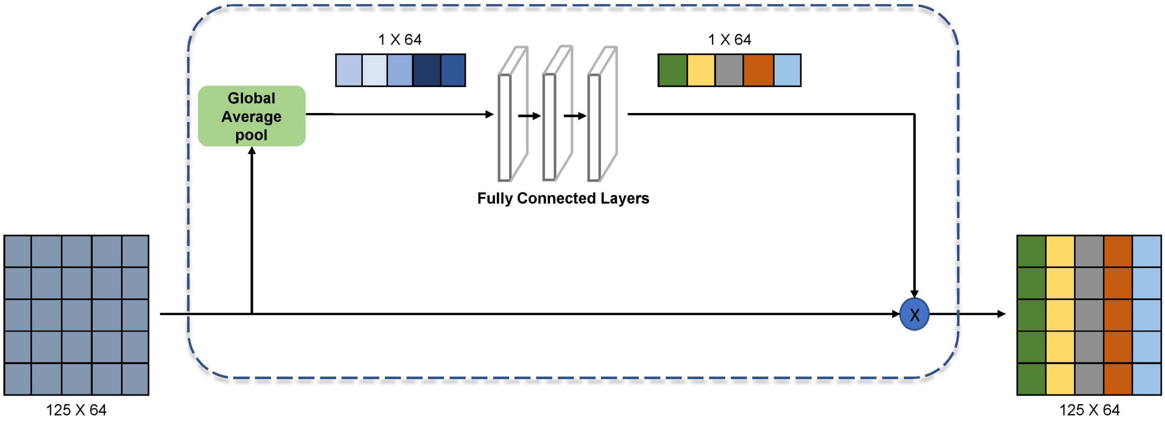

FIGURE 4. Schematic representation of the Res-SE block.

the cost of computational complexity and overfitting of the dataset.

3) RES-SE BLOCK

In this block we introduce a residual approach where an attention based architectural unit is applied in parallel to the convolutional route to model the inter dependencies between the channels of the convolutional feature map. The mechanism is called Squeeze and Excitation (SE) Attention route and it performs dynamic channel wise feature re-calibration to extract global information so that it can selectively pay more attention to the informative features and subdue others. The whole operation is completed in two steps: i) Squeeze and ii) Excitation and can be seen in Fig. 4. The different colors in the output feature map represents the various weight of attention that are put on individual channel of the input.

In Squeeze stage, the global information abstraction is performed by applying a global average pooling operation to generate an embedding of the global distribution of channel-wise feature responses. Consequently, the two-dimensional features are compressed along the spatial dimension and mapped into a one-dimensional feature vector that demonstrates the global response distribution of the overall feature map. The output feature vector is denoted as Z:

$$\mathbf{z} = \{z_c \mid c = 1, 2, 3, \ldots, C\} \in \mathbb{R}^C \tag{5}$$

Here, Z_c is the values of the feature vector for different c_{th} channel:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} \mathbf{X}_c(i, j) \tag{6}$$

In the equation, X_c is a feature map with width W for c_{th} channel that was extracted by the SCNN block in the backbone. For our model of operation channel number was fixed to 64 for optimum performance.

In the excitation section, two densely connected layers with ReLU activation function are constructed to learn nonlinear interactions between channels as well as the mutually

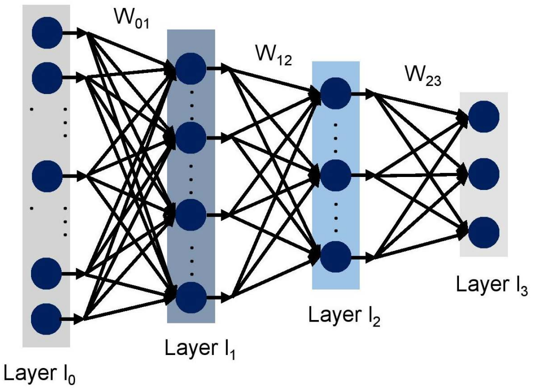

FIGURE 5. Schematic representation of the densely connected layers.

inclusive relationships. To fully capture the channel wise dependencies a self-gating mechanism with sigmoid layer was built to extract channel weights w:

$$\omega = \psi(W_2 * \sigma(W_1 * AvgPool(\mathbf{X}_{in}))) \tag{7}$$

where, ψ represents the sigmoid function and σ is the ReLU activation. Finally, the channel weights are multiplied to the conv route for improved feature selection.

4) DENSE LAYER

As demonstrated in Fig. 5, the flattened temporal feature vector extracted by the addition of the feature maps of both the Res-SEA block and the series SCNN block are directly fed into a stack of densely connected layers to converge the model towards the final prediction. The equation can be stated as:

$$l_i = \sigma(W_i l_{i-1} + b_i) \quad \forall i \in \{1, 2, 3, 4\} \tag{8}$$

Here, l_i is the output and b_i is the bias vector of the i_{th} dense layer. For our model, four dense layers constructed in series for global feature extraction demonstrated optimum performance. Finally, output vector from the last dense layer was mapped into the final prediction of hypoxemia severity using softmax activation function, whose equation is given by:

$$softmax(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{9}$$

where x represents the values from the neurons of the output layer, i is a random class for prediction and j represents the total number of classes for a given problem whose value, for our objective, was selected to be 3.

D. LOSS FUNCTION

After designing the model, the network was set to train itself using the preprocessed sampled balanced dataset. However, to optimize the training, the validation loss should be minimized and for this, the categorical cross entropy (CCE) loss function was defined so that correct severity prediction could be generated. If we consider a training set consisting of N pairs: (x_1, t_1), (x_2, t_2), (x_3, t_3), ..., (x_N, t_N), where x_i denotes the i_{th} input vector and t_i denotes the corresponding annotation target, and y_i is the model output, then the CCE loss can be defined as:

$$\mathcal{L}_{CC} = -\frac{1}{n}\sum_{i=1}^{N}\sum_{c=1}^{C}(p_{ic} log(y_{ic})) \tag{10}$$

where p_{ic} whether the i_{th} training pattern belongs to c label and output y_{ic} is the predicted probability distribution for i_{th} observation belonging to label c [25]. As for CCE loss function, the targets must be categorical, the annotation label was converted from integer to one-hot-encoded and then the whole dataset was applied to the network for model creation and validation.

III. RESULTS AND DISCUSSION

This section is divided into several parts. At first the dataset used to analyze the model performance is described. Then the evaluation metrics used in this paper is mentioned. Finally, the model is analyzed by varying the parameters and hyperparameters and compared with other deep learning and machine learning approaches.

A. DATABASE

To validate the proposed methodology a suitable database was to be selected at first. In this regard a large public physionet [26] dataset called "BIDMC PPG and Respiration Dataset" [27] from the original publication [28] was chosen for detailed analysis of the robustness of the scheme. The data was collected from several severely ill patients at the Beth Israel Deaconess Medical Centre. Two annotators were appointed to manually annotate each and individual breath in each recording utilizing impedance pneumography to derive reference respiratory rate (RR) values for the purpose of assisting RR estimation, which we would not need for the task of hypoxemia severity prediction. There are 53 recording in total, each containing PPG signal sampled at 125 Hz. The data points from the same samples correlate with each other. Each recording contains 60001 samples of data. For the model input, 1 second of frame length was chosen, thereby making 480 frames for each patient and bringing the total number of frames to 25440. The corresponding blood oxygen saturation levels (SpO2), sampled at 1 Hz, are also present in the database. While the original source of the dataset, MIMIC-II [29], recorded data for the entire stay of the patients, Pimentel et al. [28] randomly selected 8 minutes of data per patient. Our goal is to predict hypoxemia severity from just a one-second window to allow quick estimations from wearable pulse oximeters. Although there were other recordings in the dataset including electrocardiogram (ECG), heart rate (HR) and RR, we focused only on the PPG signal and utilized the corresponding SpO2 level to annotate reference hypoxemia severity based on the thresholds that we have explained in data pre-processing for its ease of collection and processed the signal afterwards to apply to the proposed deep network.

Pimentel et al. [28] mentioned that the significance of the dataset is that it demonstrates the necessity of collecting such datasets to help the scientific community improve wearable-monitoring algorithms, further aiding mobile health (m-Health) technologies, although they collected data from the hospital setting only. This gives us confidence that utilizing this dataset would help us prepare an algorithm that efficiently estimates hypoxemia severity in hospital in-patients. However, to expand into the m-Health domain, extensive study needs to be performed besides collecting a large amount of PPG data outside of hospital settings with the help of wearable devices.

The setup of the data collection procedure of the source ensures that they don't involve intermittent hypoxemia correlated with sleep disorders by including data from patients during their entire stay and not only during their sleep. However, although the hospital setting ensures that environmental factors such as low oxygen level aren't causing the hypoxemia, it's possible that temperature or other conditions might cause peripheral vasoconstriction on fingers, limiting the reliability of identifying the root cause of having a low SpO2 level in such situations or the positioning of sensors. Therefore, vasoconstriction-related limitations of pulse oximetry mentioned in [30] apply here too. Since the patients were admitted to the ICU, it is safe to assume that the randomly selected data may include effects of medications such as analgecis or sedatives as mentioned in [31]. However, Saeed et al. [29] has reported correlation between low SpO2 levels in ICU patients and their mortality rates. Therefore, it can be important to be able to quickly identify the severity of such situations.

B. EVALUATION METRICS

In this paper, various traditional metrics have been chosen for the evaluation of the proposed method such as F1 score, accuracy, precision and Cohen's Kappa score as described in the equations below:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{11}$$

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

$$F_1\ score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{14}$$

$$Cohen's\ Kappa = \frac{P_0 - P_e}{1 - P_e} \tag{15}$$

Since, different metrics have been used to analyze the performance, the experiments have been carried out in a systematic way to ensure the optimum result. In our research, different parameters and factors have been selected and modified to realize their effect on model performance.

C. PERFORMANCE EVALUATION AND COMPARISON

Initially, to demonstrate the importance of applying deep neural network for this specific objective over machine learning approach, we analyzed the performance of different machine learning classifiers such as Random forest, Naive Bayes, K-Nearest Neighbor (KNN) with different values of K and compared them to simple CNN layers and finally to our proposed method. The summary can be viewed in Table 1. Class 0, Class 1 and Class 2 in the table refer to the performance in the individual segments of original sample of Normal, Moderate and Severe Hypoxemia cases in the test set. The huge difference in per class F_1 score prediction achieved from machine learning approaches confirms the need of Deep Learning. However, as the result of simple CNN suggests, a deeper model is required to acquire better result, thereby better suited for real life application. As the high efficiency of deep learning models over machine learning can be fairly comprehended, a statistical analysis should be performed to realise the appropriate depth of the model for this particular objective. The increased performance of model with the increment of CNN layer or series CNN block in the proposed method is displayed in Table 2. For CNN layer greater than 5, results in the overfitting of the model. Layer number lesser than 5 however results in poor performance. As the need for Deep Learning has been justified for this application, the individual effect of attention block and series CNN block must be analyzed. For this purpose, the performance of individual routes and have been summarized in Table 3. Although both can fairly detect each class frames, only by merging them altogether can result in the optimum performance. The combined architecture clearly outperforms the individual performance in all the evaluation metrics, thus justifying the application of residual attention with the traditional series convolutional approach.

The number of filters may also affect the model performance. Keeping that in mind, Fig. 6 shows the values of different parameter metrics of the test set while the number

TABLE 1. Demonstration of the high performance of deep learning approach over machine learning.

Procedure	F1 score for Class			Accuracy
	Class 0	Class 1	Class 2	
Random Forest	0.95890	0.66315	0.68124	0.92689
Naïve Bayes	0.29003	0.02210	0.05677	0.18305
KNN (K = 4)	0.94294	0.33855	0.40343	0.88601
KNN (K = 5)	0.92958	0.33058	0.37620	0.86321
KNN (K = 6)	0.93416	0.33537	0.38710	0.87146
CNN (1 layer)	0.92760	0.40901	0.34286	0.85941
CNN (2 layer)	0.97529	0.71035	0.88293	0.95453
Proposed	**0.98075**	**0.75703**	**0.93734**	**0.96502**

TABLE 2. Performance analysis by varying the number of SCNN block.

SCNN Block	F1 score for Class			Accuracy	Cohen Kappa
	Class 0	Class 1	Class 2		
2	0.97350	0.69174	0.91811	0.95191	0.78988
3	0.97669	0.72640	0.90777	0.95781	0.76083
4	0.97854	0.73922	0.92346	0.96108	0.77572
5	**0.98075**	**0.75703**	0.93734	**0.96502**	**0.79427**
6	0.97832	0.73204	**0.93970**	0.96069	0.77387

TABLE 3. Performance analysis of the blocks. (Both individual and combined.)

Procedure	F1 score for Class			Accuracy	Cohen Kappa
	Class 0	Class 1	Class 2		
SCNN	0.97876	0.74506	0.91220	0.96148	0.77756
Res-SEA	0.96820	0.65018	0.81330	0.94064	0.67868
Combined	**0.98075**	**0.75703**	**0.93734**	**0.96502**	**0.79427**

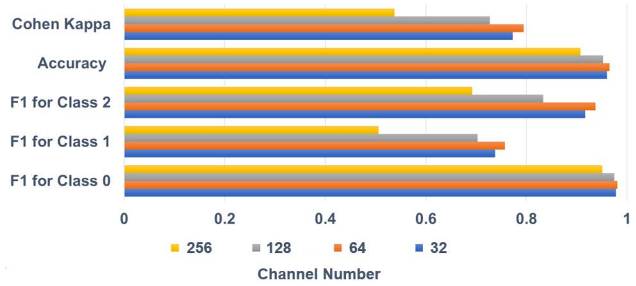

FIGURE 6. Variation of the performance metrics value with the change in channel number.

TABLE 4. Change of performance with the variation of node number in dense layers.

Node number for Dense layers	F1 score for Class			Cohen Kappa
	Class 0	Class 1	Class 2	
512-256-128-3	0.97987	**0.75779**	0.90777	0.78680
256-128-32-3	**0.98075**	0.75703	**0.93734**	**0.79427**
256-64-16-3	0.96592	0.70103	0.70746	0.68135
128-64-32-3	0.97476	0.71132	0.89904	0.74584

TABLE 5. Performance of the proposed architecture.

Metric	Class 0	Class 1	Class 2
F1 score	0.98075	0.75703	0.93734
Accuracy	0.96222	1.00	1.00
Precision	1.00	0.60905	0.88208
Overall Accuracy	0.96502		
Overall F1 Score	0.95023		
Cohen Kappa	0.79427		

TABLE 6. Effect of sampling process on the number of frames in train and test set.

Hypoxemia Severity	Sampling Process	Sample Number			
		Train Set		Validation Set	
		Before	After	Before	After
Normal	Undersampling	11692	4900	5011	2100
Moderate	Oversampling	482	5016	206	2149
Critical	Oversampling	292	4869	125	2087

TABLE 7. Performance comparison with the variation of frame length.

Class	Class 0		Class 1		Class 2	
Frame Length in Second	**1 sec**	**2 sec**	**1 sec**	**2 sec**	**1 sec**	**2 sec**
Precision	1.00	0.97	0.61	0.40	0.88	0.54
Recall	0.96	0.96	1.00	0.42	1.00	0.53
F1 Score	0.98	0.97	0.76	0.41	0.94	0.54

is varied. It can be seen that taking 64 channel results in optimum performance. Although reducing the channel number to 32 result is almost similar model performance, channel numbers more than 64 result in drastic performance degradation, especially in detecting moderate and severe hypoxemia.

While varying the CNN layers, the nodes in dense classification layers have been kept fixed. As by method of inspection, the optimum number of layers have been detected, the effect of the node number of dense layers in the model must be observed and the summary can be seen in Table 4. Here the node number of final layer is 3 for all cases to keep in accordance the 3 labels of Hypoxemia severity.

After varying the parameters and finalizing the values of individual variable, the value of different performance metrics of the proposed model can be seen in Table 5. For the final model, each CNN block has 64 filters with kernel size of 3. 4 layers have been chosen to be cascaded together in the series CNN block. While training the model, the data had been divided into 2 segments by random stratification process-70% data were chosen for the training set and 30% data were chosen for the test set respectively. For the validation dataset, samples were chosen from the last training set samples provided before shuffling, and 30% of the train data was used to generate this validation set to fine-tune the hyperparameters of the model to ensure optimum performance. It is to be noticed that the validation data was used only for evaluating the architecture, it was not used to train the model. The data in the test set did not contain any frame used in the train set either, which later underwent a series of sampling processes for data balancing. Therefore, the test data only contained new data to ensure universal performance. The effect of sampling process on the dataset can be seen in Table 6. The test set was kept isolated and did not undergo the sampling process. The model was trained for 150 epochs and gained 94.52% validation accuracy with validation loss of 0.24. The accuracy curves of the model can be seen in Fig. 7.

To compare the effect of frame length variation, the proposed 1 second frame length approach was compared to 2 second frame length approach. The complete comparison can be seen in Table 7. It can be seen that 1 second frame approach supersedes 2 second approach although the reason may well be the lower number of frames in the training and test dataset, as increasing the sample number in each frame resulted in a lower number of frames for the model to train.

To demonstrate its efficiency, the performance of the proposed model has been compared with other conventional deep networks such as Resnet, Inception Net, Google Net and VGG16-net and the result is shown in Table 8. For implementation purpose, the feature map extracted from the first CNN block was used as input to individual deep network. It can be seen from Table 6 that although the deep networks have shown better performance than the previous machine learning models, our proposed model outperforms the existing networks in almost every parameter.

Although it is true that this paper primarily focuses on accuracy and sensitivity, it has only been done to eradicate the chance of desensitization in case of emergency ICU patients. The false negative rate of a particular model can also be comprehended by the precision-recall curve, where a high recall value relates to a low false negative rate, and a high area under the curve represents both high recall and high precision. The precision-recall curve for the proposed model can be seen in Figure 8, where unlike the softmax operation, mentioned in section II.C.4, the curve applies per class binary thresholds to determine the PR values. Despite that, it approximates the precision recall trade-off. It can be seen that the model demonstrates a considerable trade-off between precision and recall.

To the best of our knowledge, there is no published work to detect the severity level of oxygen scarcity using Deep Neural Network. Although publication has been found regarding

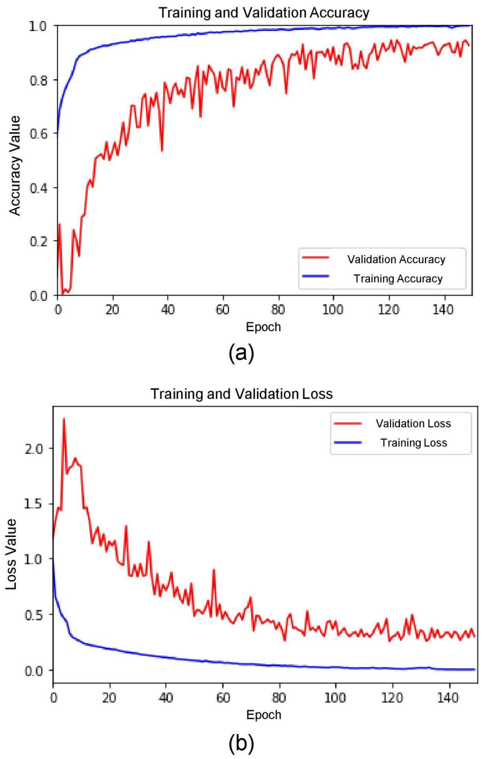

FIGURE 7. Performance of the proposed model (Epoch = 150) (a) Accuracy curve (b) Loss curve.

TABLE 8. Performance comparison among various established deep neural networks.

Procedure	F1 score for Class			Cohen Kappa
	Class 0	Class 1	Class 2	
Inception	0.96471	0.64164	0.84332	0.67310
AlexNet	0.98016	0.74876	**0.94444**	0.78923
VGG16 net	0.97454	0.69686	0.93199	0.74260
Proposed	**0.98075**	**0.75703**	0.93734	**0.79427**

TABLE 9. Performance comparison with existing approach.

Method	Metric	Class 0	Class 1	Class 2
BACDT [22]	**F1 score**	0.78	0.65	0.96
	Precision	0.80	0.67	0.95
	Recall	0.76	0.62	0.96
Proposed	**F1 score**	0.98	0.76	0.94
	Precision	1.00	0.61	0.88
	Recall	0.96	1.00	1.00

machine learning approach where bootstrap aggregation of complex decision trees (BACDT) has been applied for oxygen level prediction in different database, the performance is not stable for individual classes as can be seen from Table 9 whereas the proposed method demonstrates greater stability for all the classes. The high performance of the model by using only PPG signal as input, makes the method a promising topic to investigate and implement in near future.

IV. FUTURE PROSPECTIVE

Although the proposed method has demonstrated considerable performance in comparison to other deep neural network and existing approach, there are still some issues that need to be acknowledged and completed in near future. The issues that can be addressed are mentioned in this section.

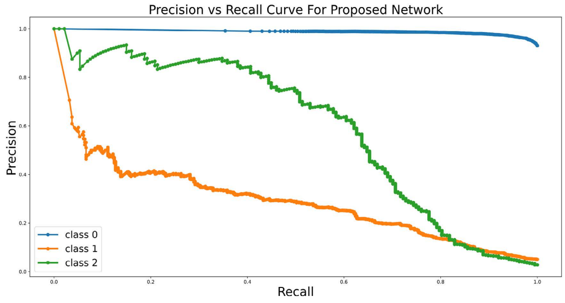

FIGURE 8. Precision-recall curve for proposed model.

A. GENERALIZATION GAP

As we can observe in Fig. 7, due to data limitations, although the oversampling-undersampling based data balancing methods are helpful based on our performance metrics, there is a noticeable gap between training and validation loss and accuracy curves, which gradually decreases as the model keeps learning, as supported by [32]. This can be related to overfitting or generalization gaps. In future, further work can be done to address this gap, inspired by methods proposed in works such as [33], [34], and [35], etc.

B. CONSIDERATION OF HEART RATE LESS THAN 60 BIT PER MINUTE (BPM)

As the frame length is taken to be 1 second in this research, there may arise a possibility when the frame will contain no heartbeat at all if the beat rate of the subject is less than 60 BPM. Therefore taking frame length of 2 second should be more appropriate approach. Yet, applying this method on the BIDMC database [26] cannot generate satisfying results due to the very low number of 2 second frames, especially for the case of class 1 and class 2, as can be seen from Table 7. We plan to utilize a larger dataset to analyze this 2 second approach and compare the performance with the variation of frame length.

C. ABSENCE OF PATIENT HOLD-OUT TESTING METHOD

Due to the frame number constraint, a mixed data approach had to be performed to analyze the model performance, where different frames of the same patient were present in training and test set. Although it was made sure that no frame existed in both set, the process does not proof the universality of the proposed method. Moreover, excluding certain number of patients' data and isolating them only for test purpose will severely affect the model training as it will not have enough unique training samples for class 1 and 2. Therefore, a larger dataset will be employed in the future for the verification of the model universality by performing the patient hold-out test.

V. CONCLUSION

In this paper, a new approach for severity prediction of Hypoxemia using PPG signal alone has been proposed. Traditional application of Pulse oximeter does demonstrate high sensitivity towards detecting oxygen degradation, yet its high rate of false alarm might lead to desensitization of the care givers. To the best of our knowledge, there has been no other research paper that has applied deep learning in predicting the saturation level. The incorporation of convolutional path and the attention route in our model has succeeded in extracting the optimum features from the input which can be easily deducted by observing the high performance of the method. Additionally, the manuscript explores the changing effect of various parameters of the model and compares the result with existing machine learning model. The high performance in all the evaluation metrics ensures the potentiality of the model for practical applications of hypoxemia severity level predictions.

VI. ACKNOWLEDGMENT

The authors would like to thank the Department of Electrical and Electronic Engineering, Bangladesh University of Engineering and Technology (BUET) for constant support.

REFERENCES

[1] C. Decaro et al., "Machine learning approach for prediction of hematic parameters in hemodialysis patients," *IEEE J. Transl. Eng. Health Med.*, vol. 7, 2019, Art. no. 4100308.

[2] P. P. Mehta et al., "Can a validated sleep apnea scoring system predict cardiopulmonary events using propofol sedation for routine EGD or colonoscopy? A prospective cohort study," *Gastrointestinal Endoscopy*, vol. 79, no. 3, pp. 436–444, 2014.

[3] T. Ahrens, K. A. R. Basham, and K. Rutherford, *Essentials of Oxygenation: Implication for Clinical Practice*. Burlington, MA, USA: Jones & Bartlett Learning, 1993.

[4] J. Bergmann et al., "356: Predicting Hypoxemia in ICU Patients," *Crit. Care Med.*, vol. 49, no. 1, p. 167, 2021.

[5] E. P. van Schaik et al., "Hypoxemia during procedural sedation in adult patients: A retrospective observational study," *Can. J. Anesthesia/J. Canadien d'Anesthésie*, vol. 68, no. 9, pp. 1349–1357, 2021.

[6] G. A. Coté et al., "Incidence of sedation-related complications with propofol use during advanced endoscopic procedures," *Clin. Gastroenterol. Hepatol.*, vol. 8, no. 2, pp. 137–142, 2010.

[7] G. D. Simpson, M. J. Ross, D. W. McKeown, and D. C. Ray, "Tracheal intubation in the critically ill: A multi-centre national study of practice and complications," *Brit. J. Anaesthesia*, vol. 108, no. 5, pp. 792–799, May 2012.

[8] S. Jaber et al., "Clinical practice and risk factors for immediate complications of endotracheal intubation in the intensive care unit: A prospective, multiple-center study," *Crit. Care Med.*, vol. 34, no. 9, pp. 2355–2361, 2006.

[9] D. E. G. Griesdale, T. L. Bosma, T. Kurth, G. Isac, and D. R. Chittock, "Complications of endotracheal intubation in the critically ill," *Intensive Care Med.*, vol. 34, no. 10, pp. 1835–1842, Oct. 2008.

[10] A. De Jong et al., "Early identification of patients at risk for difficult intubation in the intensive care unit: Development and validation of the MACOCHA score in a multicenter cohort study," *Amer. J. Respiratory Crit. Care Med.*, vol. 187, no. 8, pp. 832–839, 2013.

[11] T. C. Mort, "The incidence and risk factors for cardiac arrest during emergency tracheal intubation: A justification for incorporating the ASA guidelines in the remote location," *J. Clin. Anesthesia*, vol. 16, no. 7, pp. 508–516, Nov. 2004.

[12] A. De Jong et al., "Cardiac arrest and mortality related to intubation procedure in critically ill adult patients: A multicenter cohort study," *Crit. Care Med.*, vol. 46, no. 4, pp. 532–539, 2018.

[13] A. C. McKown et al., "Risk factors for and prediction of hypoxemia during tracheal intubation of critically ill adults," *Ann. Amer. Thoracic Soc.*, vol. 15, no. 11, pp. 1320–1327, 2018.

[14] W. Geng et al., "A prediction model for hypoxemia during routine sedation for gastrointestinal endoscopy," *Clinics*, vol. 73, p. e513, Nov. 2018.

[15] T. Y. Abay and P. A. Kyriacou, "Photoplethysmography for blood volumes and oxygenation changes during intermittent vascular occlusions," *J. Clin. Monitor. Comput.*, vol. 32, no. 3, pp. 447–455, Jun. 2018.

[16] T. Aoyagi, "Pulse oximetry: Its origin and development," in *Proc. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 7, 1992, pp. 2858–2859.

[17] B. Bohnhorst, C. S. Peter, and C. F. Poets, "Pulse oximeters' reliability in detecting hypoxemia and bradycardia: Comparison between a conventional and two new generation oximeters," *Crit. Care Med.*, vol. 28, no. 5, pp. 1565–1568, 2000.

[18] R. Sabar and E. Zmora, "Nurses' response to alarms from monitoring systems in NICU. 1027," *Pediatric Res.*, vol. 41, no. 4, p. 174, 1997.

[19] S. T. Lawless, "Crying wolf: False alarms in a pediatric intensive care unit," *Crit. Care Med.*, vol. 22, no. 6, pp. 981–985, Jun. 1994.

[20] W. Geng, H. Tang, A. Sharma, Y. Zhao, Y. Yan, and W. Hong, "An artificial neural network model for prediction of hypoxemia during sedation for gastrointestinal endoscopy," *J. Int. Med. Res.*, vol. 47, no. 5, pp. 2097–2103, May 2019.

[21] S. Ghazal, M. Sauthier, D. Brossier, W. Bouachir, P. A. Jouvet, and R. Noumeir, "Using machine learning models to predict oxygen saturation following ventilator support adjustment in critically ill children: A single center pilot study," *PLoS ONE*, vol. 14, no. 2, Feb. 2019, Art. no. e0198921.

[22] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, Jun. 2008, pp. 1322–1328.

[23] M. Zorkeflee, K. R. Ku-Mahamud, and A. Mohamed Din, "A conceptual model of enhanced undersampling technique," in *Proc. Knowl. Manag. Int. Conf. (KMICe)*, Langkawi, Malaysia, 2014. [Online]. Available: https://repo.uum.edu.my/id/eprint/13093

[24] A. L. Maas et al., "Rectifier nonlinearities improve neural network acoustic models," in *Proc. ICML*, 2013, vol. 30, no. 1, p. 3.

[25] A. Rusiecki, "Trimmed categorical cross-entropy for deep learning with label noise," *Electron. Lett.*, vol. 55, no. 6, pp. 319–320, 2019.

[26] A. L. Goldberger et al., "Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000.

[27] M. Pimentel, A. Johnson, P. Charlton, and D. Clifton. (2018). *BIDMC PPG and Respiration Dataset*. [Online]. Available: https://physionet.org/content/bidmc/1.0.0/

[28] M. A. Pimentel et al., "Toward a robust estimation of respiratory rate from pulse oximeters," *IEEE Trans. Biomed. Eng.*, vol. 64, no. 8, pp. 1914–1923, Aug. 2017.

[29] M. Saeed et al., "Multiparameter intelligent monitoring in intensive care II (MIMIC-II): A public-access intensive care unit database," *Crit. Care Med.*, vol. 39, no. 5, p. 952, 2011.

[30] P. Talke and C. Stapelfeldt, "Effect of peripheral vasoconstriction on pulse oximetry," *J. Clin. Monitor. Comput.*, vol. 20, no. 5, pp. 305–309, Oct. 2006.

[31] P. Pandharipande and McGrane, "Sedation in the intensive care setting," *Clin. Pharmacol., Adv. Appl.*, vol. 4, p. 53, Oct. 2012.

[32] E. Hoffer, I. Hubara, and D. Soudry, "Train longer, generalize better: Closing the generalization gap in large batch training of neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[33] X. Ying, "An overview of overfitting and its solutions," in *Proc. J. Phys., Conf.*, 2019, vol. 1168, no. 2, Art. no. 022022.

[34] L. Wu and Z. Zhu, "Towards understanding generalization of deep learning: Perspective of loss landscapes," 2017, *arXiv:1706.10239*.

[35] N. Shirish Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. Tak Peter Tang, "On large-batch training for deep learning: Generalization gap and sharp minima," 2016, *arXiv:1609.04836*.