Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 2 (2025)
https://www.jneonatalsurg.com

OPEN ACCESS

Proposal of Hybrid Deep Learning Algorithm for In Cabin Monitoring System

Woo-Jin Jung[1] and Won-hyuk Choi[*2]

[1]Department of Aeronautical System Engineering, Hanseo University, Taean 32158, Korea
[*2]Department of Avionics, Hanseo University, Taean 32158, Korea

Cite this paper as: Woo-Jin Jung, Won-hyuk Choi, (2025) Proposal of Hybrid Deep Learning Algorithm for In Cabin Monitoring System. *Journal of Neonatal Surgery*, 14 (2), 78-84.

ABSTRACT

An autonomous vehicle is defined as a vehicle that is capable of navigating and operating independently, without the necessity for input from a driver or passengers. The Society of Automotive Engineers (SAE) has established internationally recognised standards for the classification of autonomous driving technology, delineating the various levels of autonomy. The development of Level 4 autonomous vehicles is currently being led by leading technology companies, including Google, Nvidia, and Tesla. For vehicles classified as Level 3 or above, the autonomous system is required to fulfil the role of the driver. This necessitates the development of advanced decision-making algorithms to support high-level autonomy. To create an effective deep learning-based autonomous driving system, it is essential to have a diverse array of scenarios and a large dataset. This study introduces a novel image captioning algorithm that utilises a hybrid CNN-LSTM model to assess passenger conditions within the vehicle and generate corresponding scenarios. Furthermore, the study evaluates the suitability of the produced data for training a monitoring system through simulated environments.

Keywords: *LSTM, CNN, Deep Learning, Passenger Monitoring, Self-Driving*

1. INTRODUCTION

An autonomous vehicle is defined as a vehicle that is capable of operating without the input of a driver or passenger. The advancement of autonomous driving technology represents a significant field of inquiry in the contemporary era. The Society of Automotive Engineers (SAE) has established a set of standards that serve as the internationally accepted benchmark for autonomous driving. These standards categorize autonomous driving into six levels, from Level 0 to Level 5, delineating the technological capabilities, features, control mechanisms, and driving responsibilities associated with each level.

Recently, prominent technology enterprises, including Google, Nvidia, Tesla, and others, are engaged in active experimentation and development of Level 4 autonomous vehicles. As autonomous driving technology continues to advance, there is a growing demand and expectation for higher-level autonomous driving systems that can operate without driver intervention. In autonomous vehicles at Level 3 and above, the autonomous system must assume the role of the driver, necessitating the development of an advanced judgment system [1]. To construct such an advanced judgment system, it is necessary to develop a system that is capable of making accurate judgments based on sufficient data obtained from a variety of situations.

The design of an autonomous driving system based on deep learning necessitates the consideration of a multitude of scenarios and the availability of a substantial data set, as this enables the system to adapt and operate in an efficacious manner within real-world environments. A variety of research initiatives are currently underway, including data collection and processing, as well as simulation for learning purposes. In this study, we propose a convolutional neural network (CNN) long short-term memory (LSTM) hybrid model, which combines CNN and LSTM, a key technique in deep learning, to recognize the situation of occupants in an autonomous vehicle and generate scenarios based on the image capturing algorithm [2]. We also train the proposed algorithm through simulation and analyze its potential for generating data suitable for training a monitoring system. As autonomous driving systems become more sophisticated, it is increasingly important to be able to accurately assess the situation inside the vehicle. An occupant monitoring system that employs deep learning technology could play a significant role in this regard. In response to this need, this research aims to develop a system that can automatically analyses and process a range of situations within the vehicle.

2. CNN-LSTM HYBRID ALGORITHM

The structural synthesis of CCPGTs will be performed based on the creative design methodology process [7-8]. Fig. 3 shows the flow chart for the approach. The process consists of six steps:

2.1. CNN Algorithm

Convolutional neural networks (CNNs) are a highly effective algorithmic approach for addressing machine learning problems involving images and video data [3]. The fundamental component of a convolutional neural network (CNN) is the convolutional layer, which is responsible for effectively extracting features from an input image. This structure is optimized for the recognition of borders, contours, and various patterns in images, thereby enabling the effective extraction of useful information from high-dimensional data. Convolutional neural networks (CNNs) employ a multitude of filters to execute convolutional operations on the input image and apply a nonlinear activation function to discern intricate features. In this context, the term "filter" refers to a kernel with varying weights, which are used to extract different features. The filters are employed to identify disparate patterns within an image, including edges, colors, textures, and so forth. Subsequently, a nonlinear activation function is applied following the convolutional operation. This function introduces nonlinearity to the output of the neural network, thereby facilitating the learning of more complex data patterns. Notable nonlinear activation functions include the Rectified Linear Unit (ReLU) and the Sigmoid function. The nonlinearity of CNNs enables them to perform complex image processing that extends beyond the capabilities of simple linear transformations. Figure 1 illustrates the training process of the convolutional neural network (CNN) algorithm model, which is trained by dividing the photo data of real roads into layers [4].

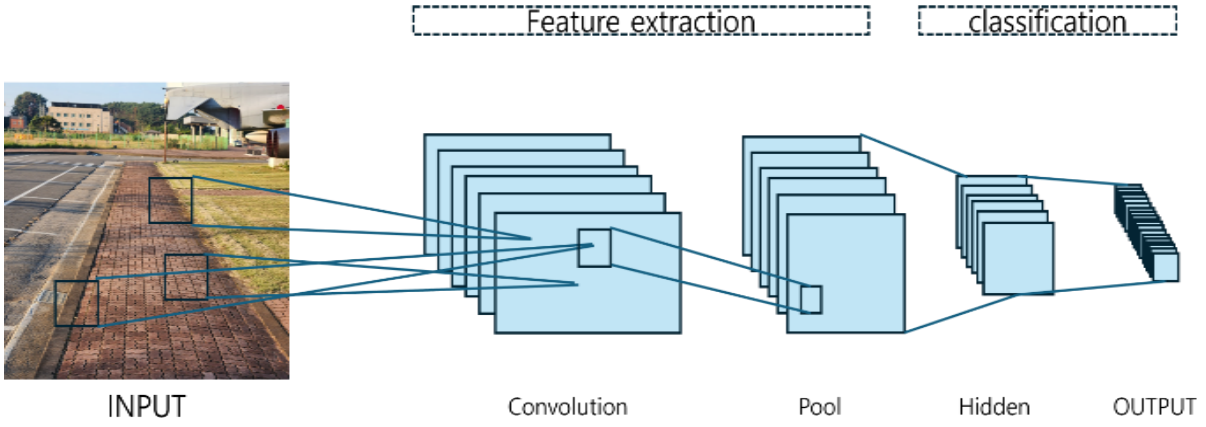

Fig. 1 CNN Algorithm

2.2. LSTM Algorithm

Long Short-Term Memory (LSTM) networks represent a specific type of Recurrent Neural Network (RNN) that has been developed to address the long-term dependency issue that arises in traditional RNNs [5]. Recurrent neural networks (RNNs) are effective at processing sequential data; however, learning becomes challenging due to the phenomenon of gradient decay, whereby crucial information diminishes over time. One solution to this problem is the Long Short-Term Memory (LSTM) algorithm. LSTMs incorporate a memory cell and gate structure that enables them to retain crucial information and discard superfluous data. LSTMs permit or prohibit the flow of information through three principal gates: the input gate, the forgetting gate, and the output gate. The input gate determines the extent to which new information is reflected in the memory cell, the forgetting gate controls the degree to which unnecessary information is erased from the cell state, and the output gate determines the way information in the cell state is output. The gates are implemented as sigmoid functions, which control the input by scaling it to a value between 0 and 1. If the value of each gate is close to 0, the information is ignored; conversely, if it is close to 1, the information is passed through. The input gates determine the extent to which new information will affect the cell state. The cell state is then updated by combining the previous state with the new input. The forget gate is responsible for determining which components of the previous cell state should be erased, which is a crucial aspect in addressing long-term dependencies. In conclusion, the output gate determines which information from the cell state is conveyed to the subsequent stage, ensuring that the network receives the requisite output information [6].

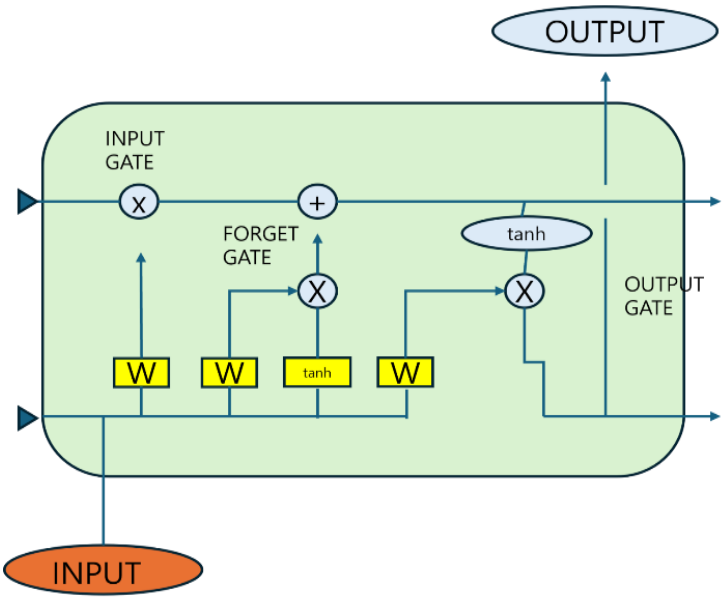

Fig. 2 LSTM Algorithm

2.3. CNN-LSTM Hybrid Algorithm Structure

The CNN-LSTM network architecture, which combines CNN and LSTM, represents a model that integrates the strengths of image processing and time series data processing in deep learning. It extracts features from images, thereby enabling predictions to be made in a manner analogous to that employed for time series data [7]. The CNN-LSTM network architecture initially performs the process of extracting salient visual information from the input image through a convolutional neural network (CNN). The principal function of a convolutional neural network (CNN) is to identify and classify distinct patterns within an image, subsequently transforming these patterns into a feature map. The filters utilized in a convolutional neural network (CNN) are designed to detect features of varying sizes, thereby facilitating the extraction of the most salient visual information from the input image. Subsequently, the image data is subjected to a dimensional reduction process through a pooling mechanism, which serves to eliminate noise and enhance computational efficiency.

The feature maps that have undergone the pooling process retain only the core information of the image, thereby enabling subsequent processing to maintain high accuracy with less data. This enables CNN to filter out the visually salient elements in the image and transform the results into vectors, which are then conveyed to the LSTM layer. Long-term and short-term memory (LSTM) is particularly adept at processing data that is presented in a sequential manner. LSTMs are frequently employed for processing time series data or natural language due to their efficacy in remembering long-term dependencies. The vectors of images extracted by the CNN are passed to the LSTM layer, where the LSTM performs processing and prediction of time series data based on the vectors. The CNN-LSTM architecture is organized in such a way that features extracted from an input image can be processed as if they were time series data.

The visual features extracted from an image are fed into the LSTM as a sequence, which allows it to model temporal dependencies. To illustrate, in the domain of image capture, a convolutional neural network (CNN) with long short-term memory (LSTM) units is employed to extract salient visual information from the input image and to generate a natural language description based on this information. In this process, the convolutional neural network (CNN) serves as an encoder, while the long short-term memory (LSTM) network performs the role of a decoder. The advantage of the CNN-LSTM structure is that it combines two key features, thereby achieving high performance in a variety of applications. While convolutional neural networks (CNNs) are highly effective at extracting local information within an image, long short-term memory (LSTM) units are well suited to learning long-term patterns and temporal dependencies. The combination of these two models is therefore highly advantageous in problems that involve simultaneous image processing and time series data. CNN-LSTM models demonstrate efficacy in applications such as image capping and video analytics. In the context of image capping, LSTMs can process the visual features of an image extracted through a CNN in order to generate a natural language description in chronological order (7).

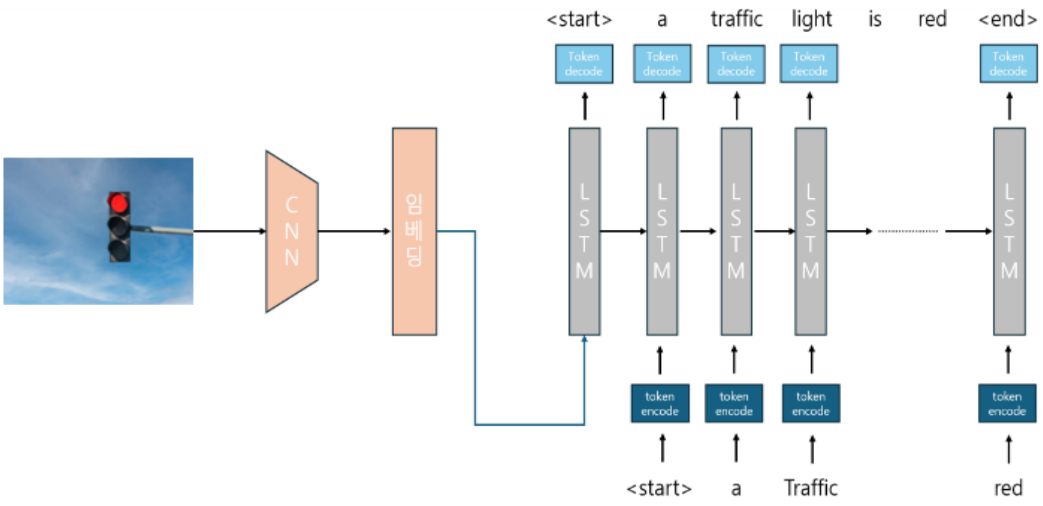

Fig. 3 CNN-LSTM Hybrid Algorithm

Figure 3 illustrates the processing of a convolutional neural network with long-short term memory (CNN-LSTM) model [8]. The figure illustrates the process of passing the given image through the convolutional neural network (CNN) model, entering the long short-term memory (LSTM) model through the embedding system of the device, and producing the desired result value based on the input string.

3. ALGORITHM ENGINEERING

3.1. Design Requirements

In the case of the CNN model, the pre-trained model ResNet 152 (Residual Network) architecture [9] was utilized. ResNet 152 is a deep learning model designed to train on very deep neural networks without a loss of performance. It was pre-trained on ImageNet, which contains over 1.4 million RGB images labeled with over 1,000 classes. The model is capable of learning effectively from deep networks through the introduction of residual blocks. The residual block incorporates skip connections that convey the input value to the subsequent layer in its original form, thereby preventing the loss of information during the learning process. Consequently, ResNet addresses the issue of gradient decay, thereby facilitating stable learning even in the context of a highly deep structure.

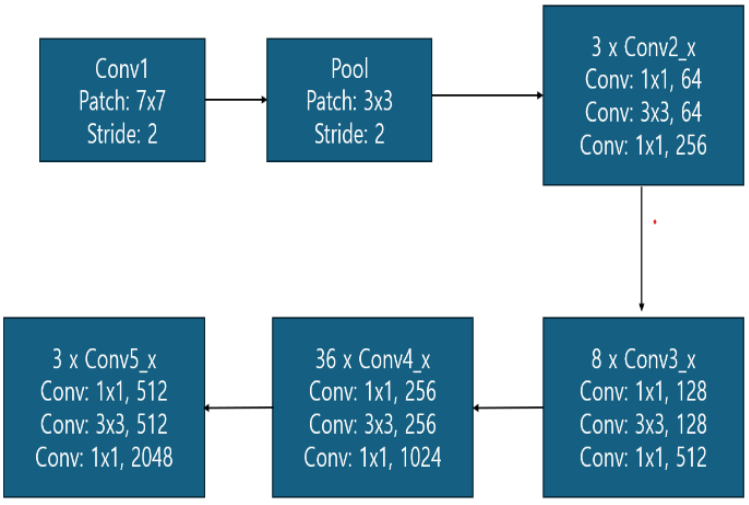

Fig. 4 ResNet152 Architecture

The ResNet-152 model comprises a total of 152 layers and has been demonstrated to perform particularly well on large datasets. The model has been pre-trained on over 1.4 million images, labelled with over 1,000 classes, including those from the ImageNet dataset. This has enabled it to learn a wide range of visual patterns and to generalize to new data. This makes it an optimal model for the complex image classification and recognition tasks employed in this thesis [10]. The model is

designed to enhance the rate of learning through residual learning and to sustain performance in deep networks without overfitting. The input data is conveyed through the skip connection in its unaltered state and is incorporated into the features extracted through the convolutional layer, thereby facilitating the acquisition of a more comprehensive representation. This effectively addresses the issue of performance degradation as the depth of the network increases [11],[13]. The final layer of ResNet-152 is removed and replaced with a fully connected layer, and a batch regularization layer is added to accelerate the learning process. This enables the model to converge more rapidly and to perform well in a variety of applications. ResNet-152 is employed in a multitude of computer vision tasks, including image classification, object detection, and image capture. Additionally, it is compatible with transfer learning utilizing pre-trained models. The loss function is cross-entropy loss, and the optimization algorithm is adaptive moment estimation (Adam) [12].

4. SIMULATION

The COCO (Common Objects in Context) 2014 train, validation, and annotation dataset was employed for the purposes of training the algorithm. Simulations were conducted on an RTX 3060 GPU environment, utilizing the PyTorch module to invoke the requisite training variables. Model checkpoints were saved at 1,000-iteration intervals throughout the fifth generation of training.

Category	Setting
GPU	RTX 3060
DATA	COCO 2014
Loss	Cross Entropy
Optimizer	Adam
Epoch	5

Fig. 5 Simulation Settings

5. DESIGN EXAMPLE

The COCO (Common Objects in Context) 2014 train, validation, and annotation dataset was employed for the purposes of training the algorithm. Simulations were conducted on an RTX 3060 GPU environment, utilizing the PyTorch module to invoke the requisite training variables. Model checkpoints were saved at 1,000-iteration intervals throughout the fifth generation of training. The training results are illustrated in Figure 5.

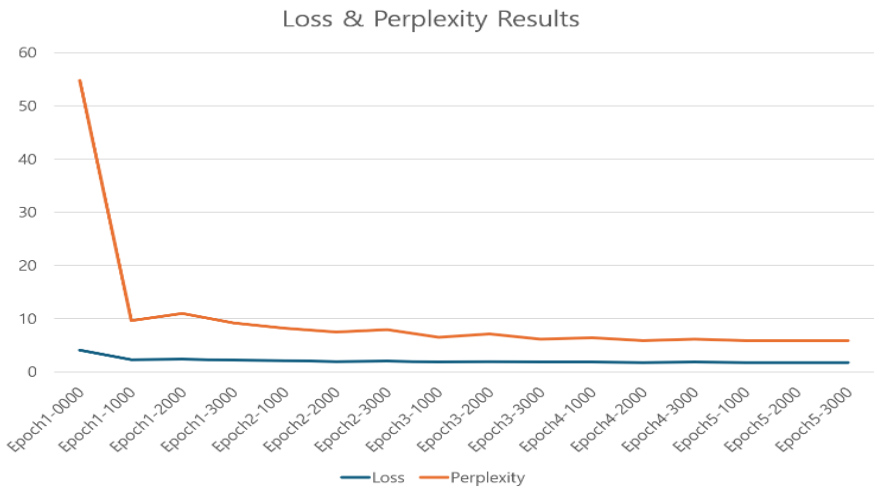

Fig. 6 Simulation Result

As illustrated in Figure 3, the values of Loss and Perplexity exhibited instability during the initial 1,000 steps. However, after this period, Loss converged to a value of 2.1, while Perplexity converged to a value of 7.6. The values of loss and perplexity resulted in minimal model confusion. The results of the test for the real-world model are presented below. It can be observed

that the model is able to correctly identify the situation. The results of executing the trained model within the Python 3 environment are presented below. It can be observed that the model produces the requisite contextual data for a scenario in which a woman is seated in a vehicle and a red traffic signal is illuminated.

Fig. 7 Model Test1

Figures 7 and 8 illustrate the outcomes of our experiments, in which we directly input the situation into the generated algorithm, which is represented as follows

Fig. 8 Model Test2

6. CONCLUSIONS

In this study, a convolutional neural network (CNN) long short-term memory (LSTM) hybrid model is proposed as a means of implementing an occupant monitoring system, which represents a crucial component of autonomous driving systems. Additionally, an image capturing algorithm is developed based on the model. To make advanced judgments in autonomous driving systems, it is essential to accurately analyze the situation inside the vehicle and learn by generating various scenarios.

The proposed model employs a convolutional neural network (CNN) to extract visual information from images and a long short-term memory (LSTM) unit to learn the dependence of time series data. This approach enables the analysis of occupant behavior within a vehicle. The simulation results demonstrate that as the training of the CNN-LSTM-based model progresses, the loss and perplexity values gradually converge towards a stable state. The results demonstrate that the proposed model is capable of functioning reliably in a multitude of scenarios. Upon feeding the trained model with images containing information about the interior and exterior of the vehicle, it was observed that the model produced outputs that were consistent

with the expected outcomes. This demonstrates that the proposed model can generate sufficient data for training the occupant monitoring system. The findings of this study suggest that the data collection and learning process can be streamlined in occupant monitoring, which is a crucial component of autonomous driving systems.

In subsequent research, the proposed model will be applied in actual autonomous driving environments to gain a more diverse understanding of situations, with the aim of enhancing the stability and reliability of autonomous driving systems. Further simulations will be conducted based on the proposed model using various data sets with the objective of further improving the performance of the model. The CNN-LSTM-based image capturing algorithm presented in this study has demonstrated that it can generate suitable data sets for training the occupant monitoring system in autonomous vehicles. This is anticipated to facilitate a more efficacious implementation of the judgment systems that are requisite for the advancement of autonomous driving systems. It is anticipated that future research will be required to enhance the safety and efficiency of autonomous driving technology through the utilization of various technological advancements.

ACKNOWLEDGMENTS

This research was supported by "Regional Innovation Strategy (RIS)" through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(MOE)(2021RIS-004)

REFERENCES

[1] Ulbrich. S, Menzel, T, Reschka. A, Schuldt. F, and Maurer. M “Defining and substantiating the terms scene, situation, and scenario for automated driving,” in 2015 IEEE 18th international conference on intelligent transportation systems, Gran Canaria, pp. 982-988, 2015.

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston: MA, pp. 3156–3164, June 2015. DOI: 10.1109/CVPR.2015.7298935.

[3] Y. LeCun et al., “Backpropagation Applied to Handwritten Zip Code Recognition,” Neural Computation, vol. 1, no. 4, pp. 541–551, 1989.

[4] KATTENBORN, Teja, et al. Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. ISPRS journal of photogrammetry and remote sensing, 2021, 173: 24-49.

[5] STAUDEMEYER, Ralf C.; MORRIS, Eric Rothstein. Understanding LSTM--a tutorial into long short-term memory recurrent neural networks. arXiv preprint arXiv:1909.09586, 2019.

[6] YU, Yong, et al. A review of recurrent neural networks: LSTM cells and network architectures. Neural computation, 2019, 31.7: 1235-1270.

[7] WANG, Jin, et al. Dimensional sentiment analysis using a regional CNN-LSTM model. In: Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers). 2016. p. 225-230.

[8] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780

[9] PANDA, Manoj Kumar, et al. Modified ResNet-152 network with hybrid pyramidal pooling for local change detection. IEEE Transactions on Artificial Intelligence, 2023.

[10] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. “ImageNet Classification with Deep Convolutional Neural Networks.”Communications of the ACM. Vol. 60, No.6, pp. 84-90, 2017.

[11] Kaming. He, Xiangyu Zhang, Shaoqing. Ren and Jian. Sun, “Deep Residaul Learning for Image Recognition,” in 2016 IEEE 18th conference on Computer Vision and Pattern Recognition(CVPR), Las Vegas, pp. 770-778, 2016.

[12] [Diederik P. Kingma, Jimmy Ba. Adam: A Method for Stochastic Optimization [Internet]. Available: https://arxiv.org/abs/1412.6980

[13] Moses, M. B., Nithya, S. E. & Parameswari, M. (2022). Internet of Things and Geographical Information System based Monitoring and Mapping of Real Time Water Quality System. International Journal of Environmental Sciences, 8(1), 27-36. https://www.theaspd.com/resources/3.%20Water%20Quality%20Monitoring%20Paper.pdf