International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING
ISSN:2147-6799 www.ijisae.org Original Research Paper

# Efficient Project Management in Construction Sites to Monitor and Track the Employees using Multi-Modal Deep Learning Model

**[1]D. Leela Dharani, [2]Manjula Pattnaik, [3]Naveen Kumar G. N, [4]Varagantham Anitha Avula, [5]Balachandra Pattanaik, [6]Shikha Maheshwari**

**Submitted**: 01/10/2023 **Revised**: 20/11/2023 **Accepted**: 01/12/2023

**Abstract**: In general, past research on automated safety monitoring using computer vision techniques has concentrated on distinct components, accounting for the safety issues individually. This is because there are a wide variety of safety issues that can arise. Recognizing the working status of construction equipment and following the movement of personnel are also instances of this kind of research. A number of researchers have come to the conclusion that it is in their best interest to implement the fundamental principle that supports the operation of a detection-based tracking system is that newly detected items either start new tracks or are mapped to existing tracks for the purpose of identification maintenance over a period of time that has been predetermined. In this paper, an Efficient project management scheme is developed using artificial intelligence. This model enables the construction sites to monitor and track the employees and this uses multi-modal deep learning (MMDL) model to track the safety of the employee. The simulation is performed with movable workers in python to test the efficacy of the MMDL and it is evaluated in terms of accuracy, precision, recall and f-measure. These performance metrics are used in the present study to check if the MMDL model is efficient in classifying the people who are working without any safety. The results show an efficient classification of instances than the other existing state-of-art models.

***Keywords:*** *Multi-Modal Deep Learning, Safety, Construction Sites, Tracking.*

## 1. Introduction

Accurate records of worker and pedestrian mobility are valuable to the professions of architecture, construction, engineering and facilities management, in addition to transportation management and emergency management. It is possible to create a working environment that is more secure and productive by gaining a better understanding of the movement patterns and behaviors of construction employees on the job site [1–7]. This is something that can be done by gaining a better understanding of the behaviors and movement patterns of construction employees.

It is possible to better understand pedestrian behavior, plan infrastructure and facility management [8, 9], plan rescue routes for first responders [10, 11], and arrange emergency evacuation by tracking pedestrians both on sidewalks and inside buildings. In addition, tracking makes it feasible to aggregate data from a large number of diverse sources and to isolate crucial features like density, flow, and speed [12].

Recent advancements in deep learning have made it possible to carry out automated data processing and have permitted its application in a wide range of academic subjects. These breakthroughs have also made it possible for deep learning to be used. The events that led to these changes took place over the course of the last few years. Detecting non-certified work [13], detecting, tracking, and analyzing the activities of earth-moving machinery [14], detecting unsafe behavior [15], detecting workers [17], and detecting untrained workers in the use of machinery [18] are all examples of applications that make use of deep learning.

This research will provide a three-part architecture with the goal of enhancing the tracking of multiple persons at the same time. This involves monitoring, detection and tracking using YOLO, CNN and tracking framework.

## 2. Literature Survey

An evaluation of the prior research on the subject of safety monitoring at construction sites has been carried out. Specifically [18] classified computer vision-based building safety monitoring strategies into three distinct

*[1]Assistant Professor, Department of Information Technology, P.V.P Siddhartha Institute of Technology, Vijayawada, Andhra Pradesh, India.*
*[2]Associate Professor, College of Business Administration, Princess Nourah Bint Abdulrahman University, Riyadh, KSA.*
*[3]Associate Professor, Department of Electronics and Communication Engineering, CMR Institute of Technology, Bengaluru, Karnataka, India.*
*[4]Department of Computer Science, Institute of Technology, Hawassa University, Awassa, Ethiopia.*
*[5]Professor, Department of Electrical and computer engineering, College of Engineering and Technology, Wallaga University, Nekemte, Ethiopia, Africa.*
*[6]Associate Professor, Directorate of Online Education, Manipal University Jaipur, Rajasthan, India.*
*[1]dharanidonepudi@gmail.com, [2]drmanjula23@gmail.com, [3]gnnaveen08@gmail.com, [4]anithav@hu.edu.et, [5]balapk1971@gmail.com, [6]shikha.maheshwari@jaipur.manipal.edu*

groups. The first category is called worker recognition, and it refers to the ability to identify physical workers.

The authors in [19] was able to successfully segment a wide variety of components on images, including people, tools, and supplies, among other things. Recent research conducted in [20] utilized a more efficient R-CNN architecture that is based on regions in order to identify employees working on scaffolding. A complicated algorithm developed by CNN was then applied to the data in order to determine the percentage of employees who are actually wearing their seatbelts while they are on the job. It was determined that those who weren't properly tethered had a lower chance of surviving a fall from a substantial height.

This was due to the fact that their chances of survival decreased. After putting deep learning into practice, there was a noticeable increase in the level of accuracy reached by detection [6]. This was due to the fact that deep learning allows for more complex patterns to be recognized. The second category includes the processes that are used to monitor the progression of different items [21].

In [22], a detection-based tracking model was established. By utilizing the SIFT algorithm to extract visual characteristics from the images. The authors in [23] were able to differentiate between humans and computers in the images. The data were inputted into a Kalman filter, which was used to make forecasts about the course that the system will take in the future based on the measurements that had already been done. These projections were based on the fact that the system had been measured previously.

In the meantime, HOG features were utilized in order to detect workers and machines in a different study [24], and a particle filter was utilized in order to follow the activities of these individuals. Despite the fact that a variety of tracking algorithms that are dependent on detection have been developed for the purpose of monitoring construction sites, the potential that deep learning holds in this area has not yet been thoroughly investigated. The third category of talents encompasses the capability of identifying activities as one of its components.

In [25], the authors to identify potentially dangerous activities in employees. A few examples of these behaviors are climbing ladders while carrying heavy goods in their hands, climbing ladders while facing the wrong direction, and reaching too far. It was possible to effectively identify potential dangers to workers, but the method that was used only recorded incidents that involved a single worker; multi-labour analysis was not taken into consideration.

The spatial-temporal interaction between workers and machinery is rarely evaluated in a holistic manner, which implies that potential effects from a range of viewpoints are not taken into account. This can have negative health consequences for workers. In order to support autonomous and real-time monitoring of on-site safety, a robust strategy that evaluates the spatial-temporal interaction between employees and equipment is necessary. It is necessary to put into practice this strategy.

## 3. Methods

We demonstrate a system that, with the assistance of deep learning, is able to monitor a building site, identify any potential dangers that may be present there, and follow the path that an worker takes as it moves. Discovering the locations of the safety authorities and keeping track of the activities they are engaged in are also essential components of this technique. In this part of the article, we will go through the essential concepts that underlie the approach.

The YOLOv3 network structure is particularly effective because it solves the challenge of target identification by recasting it as a regression problem. This allows the network to function more effectively. Because of this, finding a solution to the problem is going to be much simpler. In several different locations, the bounding box and the category to which the image belongs are both directly returned when searching for a certain image. Because of the way its layout was designed, YOLOv3 has a detection speed that is quite high, and it is capable of satisfactorily meeting the needs of real-time operations. In the following, we will discuss the algorithm in greater detail, concentrating on the multiple components that make up the algorithm, and we will analyze each of those components.

When looking into possible algorithms for supervised learning, YOLOv3 is an excellent choice to take into consideration. The very first thing that we do with a image is to grid it into squares of size SxS, and only after that do we move on to the following phase. If the center of an worker can be located within this grid, then it can be held accountable for producing an accurate forecast regarding the worker. Inside of each grid, we draw conclusions concerning the B-boxes and the levels of confidence that are connected to those B-boxes. Using this method, it is much simpler to differentiate between things that are both compact and overlapping with one another at the same time.

A confidence value (x, y,w, h,C), width (w), height (h), and the center coordinates (x, y) are the components that make up a bounding box. The bounding box is constructed out of these five individual bits of data (C). The extent to which it is thought that the worker can be contained within the bounding box that surrounds it is a good indicator of

how well it can do so. The following is the correct method for computing the figures:

$$C = Pr(worker) \times IoUtruth\ pred\ (1)$$

where

where the value of Pr(worker) shows whether or not the object in question is contained within the grid, then Pr(worker) = 1; otherwise, Pr(worker) = 0. Pr (worker) will equal 1 if the item can be discovered within the bounding box; however, if it cannot be found, it will equal 0 instead.

If the IoU value is larger than zero, this indicates that the bounding box of the worker has all of its sides. If the value is less than zero, this indicates that one or more of the sides is missing.

$$IoU = [A(GT) \cap A(PB)] / [A(GT) \cup A(PB)]\ (2)$$

where,

A - Area

GT - ground truth

PT - prediction box

The following is the level of certainty that corresponds to c, which is the highest possible level:

$$c = Pr(class(i)|worker) \times Pr(worker) \times Predicted(IoU)$$

$$= Pr(clsassi) \times Predicted(IoU)\ (3)$$

Figure 1 demonstrates that in order to carry out the process of feature extraction, it relied on the darknet53 network. The possibility of the model converging has grown ever more likely ever since the residual unit became a component of this network.

**MMDL Framework**

In this study, we analyze whether or not it is possible to recognize images by employing a CNN architecture that includes two streams by incorporating data from a wide variety of modalities in a way that is both effective and efficient. Specifically, we look at the results of our investigation. As an alternative to picking a single input pixel, which is centered on the pixel that is selected as input, and this neighborhood is then used to collect spatial information from the pixels that are next to it.

Modern CNN layouts are built from a number of layers that are placed atop one another and connected to one another. Each layer is responsible for a different function. The following is a selection of the layers that are studied within the confines of the scope of this work:

- Convolution layer: This is the initial layer.
- Activation layer: The activation layers are the ones that are in charge of the execution of non-linear operations, which are what make it possible for the networks to approximate any function. One such illustration is provided by the ReLU activation, which makes use of a thresholding operator that is not non-negative. One more illustration of this type is the sigmoid activation, which limits the output [0, 1] and is typically implemented as a predictor in the final layer of the network. These two activations are simply two different manifestations of the same phenomenon.
- Batch normalization layer: A layer that facilitates the normalization of data batches during the process. As a direct consequence of the incorporation of this layer into the architecture of the network, a normalization phase has been included. At this stage, it is ensured that all trainable layers receive inputs that are uniform with regard to the degree of uniformity across all of the characteristics. This assures that learning will take place at a rapid pace while also preserving the operation of the network.
- Pooling layer: The primary functions of a pooling layer are to lessen the dimensionality of feature maps and to make the maps invariant to inputs whose values are subject to some degree of variation. Max Pooling is the alternative that sees the most use, despite the fact that it does not permit the negative element reduction, the smoothing of gradients across the network.
- Dropout layer: The technique known as dropout is one of the most widespread and widely used methods for preventing overfitting. Dropout works by momentarily lowering the total number of parameters utilized by the network.
- Fully connected layer: In order for a layer connected to create accurate predictions, it must first digest the input characteristics and then use those digested features as inputs to subsequent layers. Only then can the layer be considered fully linked.

In this article, we conduct an analysis of a network architecture that uses a series of convolutional and pooling layers for the purpose of feature extraction, followed by dropout and batch normalization layers in order to prevent overfitting. Our goal is to improve the accuracy of the network by reducing the likelihood that it will become overly accurate. The process of feature extraction is carried out by combining all of these layers together. Following that is the final layer that uses the softmax scoring method, which is followed by two layers that are totally connected. Early fusion is a technique that we used to construct a DNN that uses memory more effectively. This was accomplished by combining data at an earlier stage in the learning process. Before the first layer that

was entirely joined together, this strategy was put into action. Specifically, the multi-modal CNN that has been demonstrated consists of three distinct modules all working independently of one another.

In order to produce features, both the first and second modules apply three-dimensional convolutional layers to the patches that they receive as input. Because our hyperspectral data consists of spatial information that can't be accurately represented by 2D models, we decided to go with the more trustworthy 3D convolution method instead.

This is to ensure that appropriate generalization is gained during training. In order to get started, the HSI image and the four individual RGB tiles needed to be combined and reduced such that they corresponded to the same spatial parameters as the ground truth data. Only then could the process begin.

Blocks of each of the training images that were exactly 25x25 pixels in size were cut out by dragging a selection tool across the screen. In order to get a greater quantity of information, the patches that are created will share their pixel space with both of the images. In addition, we selected 400 unique patches at random from each image by utilizing a single pixel as the seed for the randomization process. This resulted in a total of eight thousand (8,000) patches being produced for us to analyze across all of the various categories. Figure 1 demonstrates the proposed processing pipeline, while Figure 2 illustrates the network fine-grained parameters. Both figures are located in the same file. Both numbers are listed down here for your convenience.


**Fig 1.** Pipeline for MMDL

| **Conv 3D + ReLU (2)** |
|---|
| **Conv 3D + ReLU (4)** |
| **Max-Pooling** |
| **Batch Normalization** |
| **Conv 3D + ReLU (8)** |
| **Conv 3D + ReLU (8)** |
| **Max-Pooling** |
| **Batch Normalization** |
| **Conv 3D + ReLU (16)** |
| **Conv 3D + ReLU (16)** |
| **Max-Pooling** |
| **Batch Normalization** |
| **Conv 3D + ReLU (32)** |
| **Conv 3D + ReLU (64)** |
| **Max-Pooling** |
| **Batch Normalization** |
| **Conv 3D + ReLU (128)** |
| **Conv 3D + ReLU (256)** |
| **Conv 3D + ReLU (512)** |

**Fig 2.** Parameters of MMDL

**Distributed CNN**

The processing pipeline that would be utilized by the proposed multimodal CNN is depicted in Figure 1. One component of the network architecture (shown on the left), which processes the RGB data, and another component of the network architecture (shown on the

right), which processes the HSI data, are responsible for producing the categorization results respectively. These two constituents collaborate to form the network as a whole.

Figure 2 presents the parameters that are recommended for implementation in a multi-modal classification network. Conv3D will first carry out a three-dimensional convolution using the number of filters that the study provides, and it will then apply a ReLU non-linear activation to the output of this operation. When a layer uses the required number of filters and is then activated by the ReLU function, it is referred to as a dense layer.

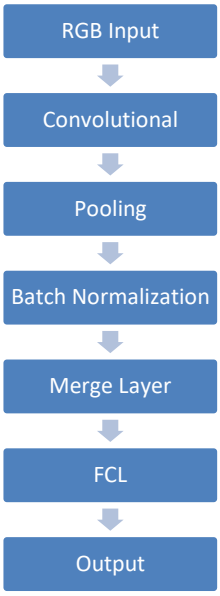


**Fig 3:** Baseline Architecture

The concept of data parallelism, as can be shown in Figure 3, provides all employees with access to the same data. On the other hand, the model of parallelism that is offered provides employees with access to a variety of data. We investigate the differences in how long the processing takes and how accurate the results are in order to put a numerical value on the performance advantages that are connected with each strategy. This allows us to put a numerical value on the performance advantages that are connected with each strategy.

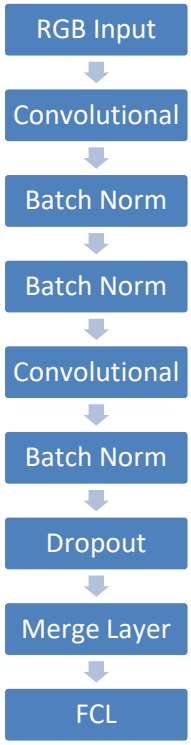


**Fig 4.** Optimized Architecture

Two distinct designs for distributed deep learning are currently being contemplated by our team as a means of optimizing the utilization of the resources that are already at our disposal within the cluster. The multimodal CNN serves as the foundation for both of these architectural designs. Because of this, the CNNs that we employ include three forks: one for each modality, in addition to a fusion and classification fork. These three forks allow the CNNs to effectively process the data that they are given. Figure 4 demonstrates that first, we built a core architecture in which each branch was delegated to a specific worker. This design may be seen in its entirety here.

Figure 4 demonstrates that the proposed distributed multimodal CNN has a baseline architecture that is made up of the following three parts: On the right, we can see the top branch of the algorithm, the bottom the HIS algorithm, and the merging component that, in the end, produces a categorization result. On the left, we can see the merging component that produces a categorization result.

When we were planning out the second architecture, we made sure to take into account how active our network generally is. Taking this into consideration allowed us to plan more effectively. As a direct result of this, we were forced to make certain modifications to the distribution of our earlier model in order to develop the one that is shown in Figure 4. We were able to successfully transmit the last block of the convolutional pooling procedure as well as the flat and dense layers of the RGB of our earlier model in order to get to the one that is shown in Figure 4. We were able to properly transmit the final block of the convolutional pooling process by applying this method. In the later stages, the software and hardware were consistently used even if they had been used previously.

**Worker Tracking Mechanism**

The observing and following pedestrians will be covered in this essay. Both the Kalman filter and the Hungarian algorithm are put to use in the process of developing the pedestrian-tracking system that our organization puts into place. The Kalman filter is responsible for making forecasts about where the pedestrian.

A comparison is done between the position of the pedestrian and the position of the pedestrian in the present frame. For the purpose of accurately representing the location prediction mechanism, we make use of a model of a process system with discrete control. One is able to model the system in the following manner by making use of a equation:

$$X(n) = W(k) + (X(k-1) \times A) + (U(k) \times B) \quad (4)$$

The value that was determined to be appropriate for this system may be written as:

$$Z(k) = V(k) + (X(k) \times H) \quad (5)$$

where

X(k) - system state, and

U(k) - system control.

Z(k) - measured value, and

H – parameter,

W(k) and V(k) - white Gaussian noise.

*A* and *B* - system parameters and

R - covariance.

The workflow of the filter can be divided up into: the prediction and the updating phase. Assuming that the current state of the system is X(k), we can use the model of the system to make a prediction about the next state of the system based on the past states. This prediction can be made in the following manner:

$$X(n|n-1) = A \times X(n-1|n-1) + B \times U(n) + W(n) \quad (6)$$

where

X (n|n − 1) - prediction of prior state, and

X(n −1|n −1) - previous state.

The covariance is hence calculated as below:

$$X(n|n-1).\ P(n|n-1) = A \times P(n-1|n-1) \times A\ T + Q \quad (7)$$

The weighting matrix is estimated using the Kalman filter as below:

$$Ng(n) = P(n|n-1) \times H\ T\ /(H \times P(n|n-1) \times H\ T + R) \quad (8)$$

Estimate $X(n|n)$ using Kalman gain as below:

$$X(n|n) = X(n|n-1) + Ng(n) \times (Z(n) - H \times X(n|n-1)) \quad (9)$$

The covariance is updated as below:

$$X(n|n).\ P(n|n) = (I - Ng(n) \times H) \times P(n|n-1) \quad (10)$$

where I - identity matrix.

When analyzing a particular IoUtrack, the predicted position of an area = actual position of the area is applied. The mismatch between the positions that were forecasted to be filled and those that were actually occupied:

$$IoUtrack = [A(PP) \cap A(RP)] / [A(PP) \cup A(RP)]$$

where

PP - predicted position

RP - real position

We entered the IoUtrack calculated into a matrix and then ran it through the Hungarian matching algorithm so that we could determine the precise position where the two frames coincide with one another. The subsequent paragraphs will provide an outline of the particular actions that need to be carried out. We will assume that there will be no loss of detection between the two cameras and will use a total of four people.

The current frame position is indicated by the symbol Wi, where {Wi , i = 1, 2, 3, 4}, and the forecast that was generated by the Kalman filter is represented by the symbol Ji, where {Ji, i = 1, 2, 3, 4}, respectively. The results of the computations performed for IoUtrack have been written down and recorded in this matrix. Each value has been multiplied by 100 so that the computations can be performed in a clear manner.

To begin, we take the value in the matrix and remove 100 from it. This is done so that we may maximize the IoUtrack between the expected location and the actual place. After that, the Hungarian matching procedure is executed, which identifies the most optimal solution to the assignment problem that may possibly be found. Two, the row of the matrix that has the lowest possible value is removed from further consideration for all other rows in the matrix. During the last step of the process, the total value of each column in the matrix will have the minimum value of each column deducted from it. Because of this, there will be exactly one zero in each row and column after this process is complete. During the final stage of the process, we will search for all of the workers that have a value of zero and work to find a way to accommodate them while using the fewest number of rows and columns that is possible.

After some investigation, it was found that the values for W2 and W4 in the corresponding row, as well as the value for J3 in the corresponding column, are all equal to 0. The goal of the fifth stage is to locate the fewest number of possible values in the rows and columns that were skipped in the previous four stages. After that, the minimal number is subtracted from all of the open numbers, and the minimum number is added at the intersection of the rows and columns in the fourth stage of the process.

The study can ensure that the zeros are distributed throughout the matrix in the most optimal manner by repeating steps 4 and 5 as many times to include all zeros is equal to the size of the matrix. If the study do this, the study will have achieved the goal of ensuring that the zeros are distributed in the most optimal manner. In the sixth step, the study will look for any occurrences of the number zero that may be present in rows and columns that do not correspond with one another. These spots in the initial allocation matrix are the ones that have the highest total number of IoUtracks assigned to them. We are able to track a pedestrian by locating them in three different frames, identifying them in two of those frames, and then doing it again in the third frame using the procedure that was just stated.

## 4. Results and Discussions

In this section, the proposed model is examined and validated. The hardware consisted of two GPUs for the purposes of building and testing the model. The dataset that was offered for the ICCV'17 PoseTrack Challenge is used. This dataset is very challenging to analyze due to the frequent occurrence of occlusion as well as the presence of dense human targets across a wide variety of image formats. The dataset contains a total of 350 instructional films, including 50 validation movies, and 300 training videos. The simulation parameters is shown in Table 1.

**Table 1:** Simulation Parameters

| Video Parameter | Value |
|---|---|
| Video | .mp4 |
| Length | 20 minutes |
| Resolution | 720p |
| Total images after video frame extraction | 1240 |

These images were taken from a wide variety of recordings that were taken by surveillance cameras that were placed at a number of different building sites. Each of the 2410 images has been annotated with bounding boxes and ground truth labels, while the images themselves depict five distinct categories of construction equipment in addition to construction people.

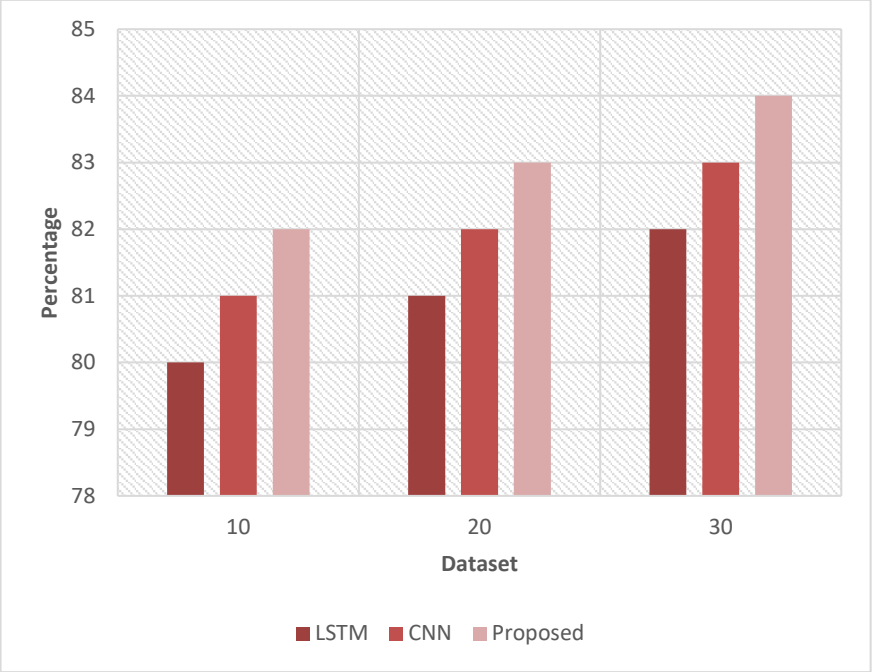


**Fig 5:** Training Accuracy

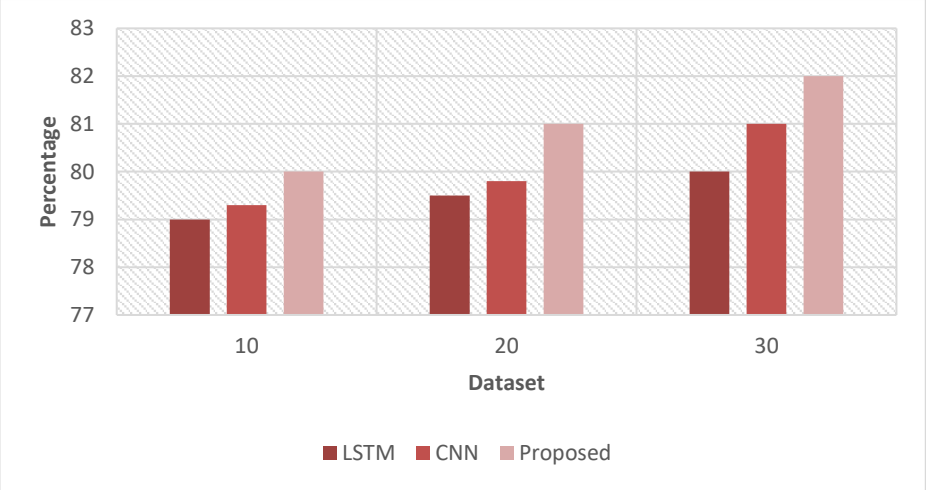


**Fig 6:** Testing Accuracy

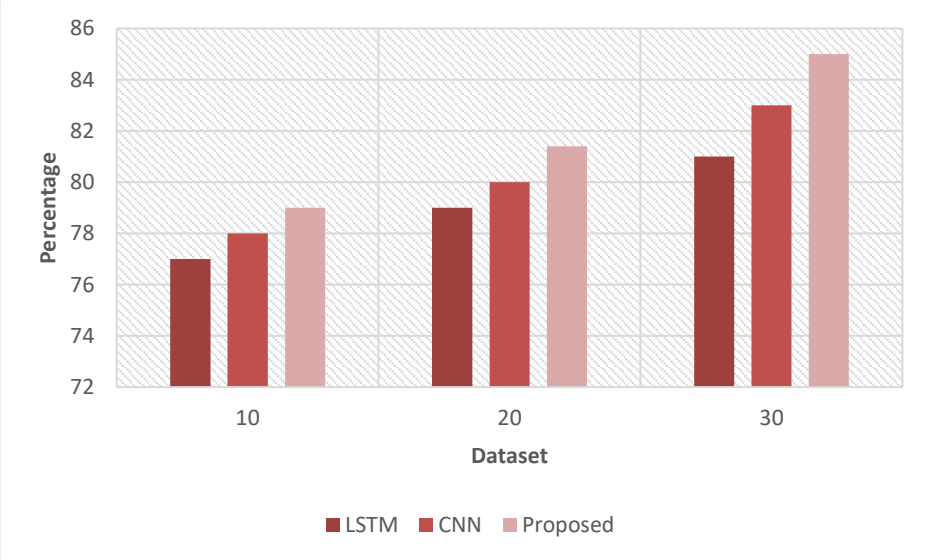


**Fig 7:** Sensitivity

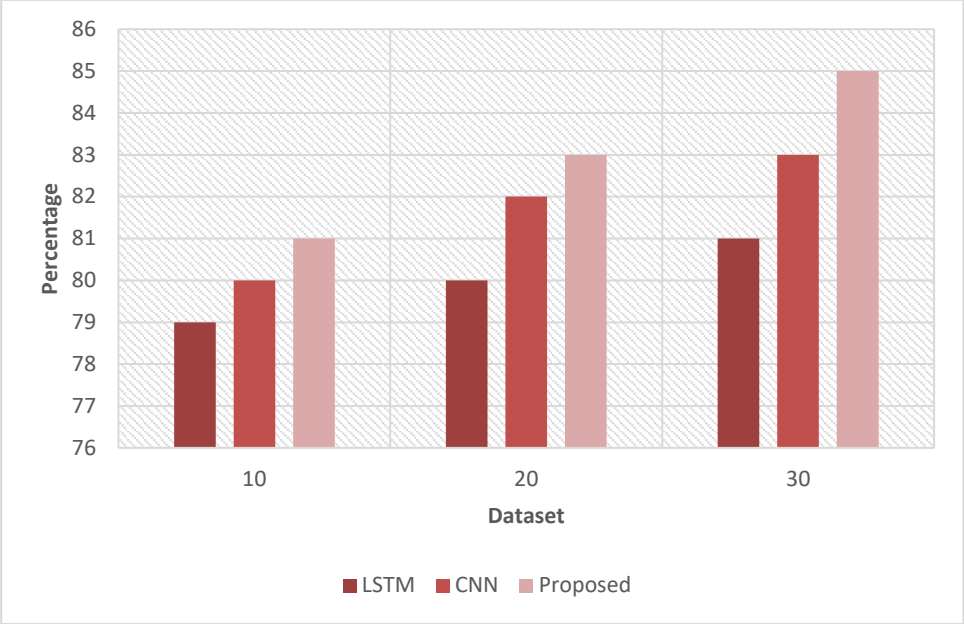


**Fig 8:** Specificity

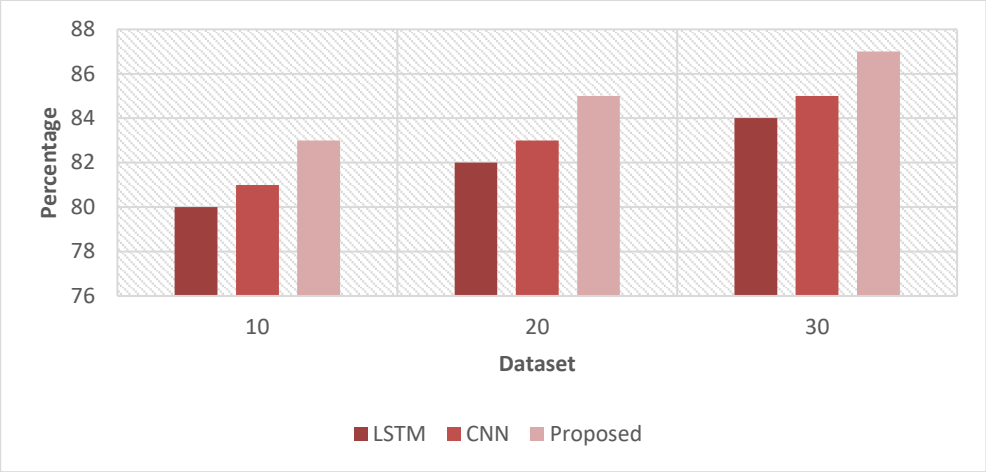


**Fig 9:** F-measure

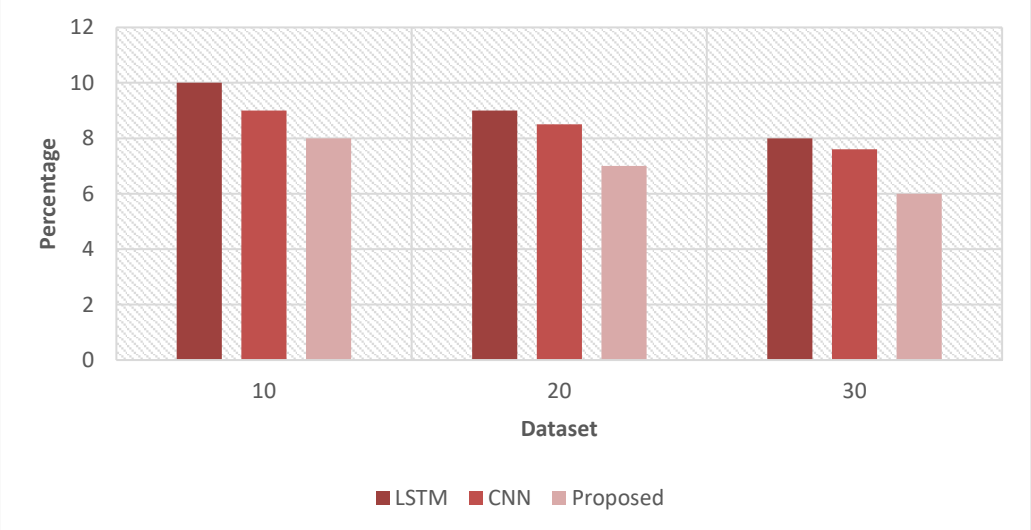


**Fig 10:** MAE

Figure 5-10 shows the results of the tracking training and testing accuracy, sensitivity, specificity, f-measure and mean absolute error (MAE) and the results shows an improved tracking of workers than the existing methods.

The vast majority of the images, which make up 90% of the total, are utilized for the purposes of model training and validation, whereas just one tenth of the images are utilized for the testing of the models in their actual environments. In conclusion, the method is illustrated by applying it to brand-new picture of a building site, which enables real-time forecasting of potential threats.

The accuracy with which worker safety statuses are assigned has been increased to a level of 88% accuracy so that the appropriate safety danger notifications may be provided. The videos obtained from the surveillance cameras that were installed at the building site are used as the basis for the experiments that are carried out. The purpose of these tests is to provide evidence that the method that has been suggested is successful. This was done in order to protect workers from receiving false warnings. Finding people working and machinery operating on building sites allowed this to be accomplished.

## 5. Conclusions

In this paper, it is seen that the construction industry is the most unsafe of all businesses because of the huge number of incidents that take place on work sites. Construction sites are incredibly dynamic places. This is due to the constant activity of individuals working at the site as well as the broad array of construction machinery that is used. One of the key elements that may lead to the emergence of dangers on the job site is the interaction that occurs between personnel and the machinery that they are operating. For this reason, it is absolutely necessary to keep a close eye on the operational condition of both the construction workers and the equipment they use, as well as to conduct an investigation into the connections that exist between these elements in terms of both space and time, in order to eliminate any potential risks.

The results of the experiment suggest that the integrated method presented for forecasting safety concerns among construction people and equipment performs effectively, but with a number of significant qualifications and constraints. We will endeavor to improve the performance of the method that has been suggested, as well as investigate other feasible options, so that we can obtain danger zones around construction equipment with greater convenience. This will allow us to obtain danger zones around construction equipment more easily.

## References


[1] Kasa, K., Burns, D., Goldenberg, M. G., Selim, O., Whyne, C., & Hardisty, M. (2022). Multi-Modal Deep Learning for Assessing Surgeon Technical Skill. *Sensors*, *22*(19), 7328.

[2] Bai, N., Nourian, P., Luo, R., & Pereira Roders, A. (2022). Heri-graphs: a dataset creation framework for multi-modal machine learning on graphs of heritage values and attributes with social media. *ISPRS International Journal of Geo-Information*, *11*(9), 469.

[3] Hofmann, S. M., Beyer, F., Lapuschkin, S., Goltermann, O., Loeffler, M., Müller, K. R., ... & Witte, A. V. (2022). Towards the interpretability of deep learning models for multi-modal neuroimaging: Finding structural changes of the ageing brain. *NeuroImage*, *261*, 119504.

[4] Liu, J., Luo, H., & Liu, H. (2022). Deep learning-based data analytics for safety in construction. *Automation in Construction*, *140*, 104302.

[5] Thiam, P., Hihn, H., Braun, D. A., Kestler, H. A., & Schwenker, F. (2021). Multi-modal pain intensity assessment based on physiological signals: A deep learning perspective. *Frontiers in Physiology*, *12*, 720464.

[6] Ahmad, Z., Jindal, R., Mukuntha, N. S., Ekbal, A., & Bhattachharyya, P. (2022). Multi-modality helps in crisis management: An attention-based deep learning approach of leveraging text for image classification. *Expert Systems with Applications*, *195*, 116626.

[7] Tan, T., Das, B., Soni, R., Fejes, M., Yang, H., Ranjan, S., ... & Avinash, G. (2022). Multi-modal trained artificial intelligence solution to triage chest X-ray for COVID-19 using pristine ground-truth, versus radiologists. *Neurocomputing*, *485*, 36-46.

[8] Zhang, W., Wu, Y., Yang, B., Hu, S., Wu, L., & Dhelim, S. (2021, August). Overview of multi-modal brain tumor mr image segmentation. In *Healthcare* (Vol. 9, No. 8, p. 1051). MDPI.

[9] Ali, S., Li, J., Pei, Y., Khurram, R., & Mahmood, T. (2022). A Comprehensive Survey on Brain Tumor Diagnosis Using Deep Learning and Emerging Hybrid Techniques with Multi-modal MR Image. *Archives of Computational Methods in Engineering*, 1-26.

[10] Cherif, E., Hell, M., & Brandmeier, M. (2022). DeepForest: Novel Deep Learning Models for Land Use and Land Cover Classification Using Multi-Temporal and-Modal Sentinel Data of the Amazon Basin. *Remote Sensing*, *14*(19), 5000.

[11] Chi, N., Wang, X., Yu, Y., Wu, M., & Yu, J. (2022). Neuronal Apoptosis in Patients with Liver Cirrhosis and Neuronal Epileptiform Discharge Model Based

upon Multi-Modal Fusion Deep Learning. *Journal of Healthcare Engineering*, *2022*.

[12] Chen, Y., Zhu, L., & Karki, D. (2022). Data Mining of Swimming Competition Technical Action Based on Machine Learning Algorithm. In *International Conference on Multi-modal Information Analytics* (pp. 570-577). Springer, Cham.

[13] Nandi, A., & Xhafa, F. (2022). A federated learning method for real-time emotion state classification from multi-modal streaming. *Methods*.

[14] Lamichhane, B., Jayasekera, D., Jakes, R., Glasser, M. F., Zhang, J., Yang, C., ... & Hawasli, A. H. (2021). Multi-modal biomarkers of low back pain: A machine learning approach. *NeuroImage: Clinical*, *29*, 102530.

[15] Kustowski, B., Gaffney, J. A., Spears, B. K., Anderson, G. J., Anirudh, R., Bremer, P. T., ... & Nora, R. C. (2022). Suppressing simulation bias in multi-modal data using transfer learning. *Machine Learning: Science and Technology*, *3*(1), 015035.

[16] Guiney, R., Santucci, E., Valman, S., Booth, A., Birley, A., Haynes, I., ... & Mills, J. (2021). Integration and analysis of multi-modal geospatial secondary data to inform management of at-risk archaeological sites. *ISPRS International Journal of Geo-Information*, *10*(9), 575.

[17] Zhang, W., Wu, Y., Yang, B., Hu, S., Wu, L., & Dhelim, S. (2021). Overview of Multi-Modal Brain Tumor MR Image Segmentation. Healthcare 2021, 9, 1051.

[18] Chai, Y., Zhou, Y., Li, W., & Jiang, Y. (2021). An explainable multi-modal hierarchical attention model for developing phishing threat intelligence. *IEEE Transactions on Dependable and Secure Computing*, *19*(2), 790-803.

[19] Bhowmik, R. T., & Most, S. P. (2022). A Personalized Respiratory Disease Exacerbation Prediction Technique Based on a Novel Spatio-Temporal Machine Learning Architecture and Local Environmental Sensor Networks. *Electronics*, *11*(16), 2562.

[20] Dai, Y., Song, Y., Liu, W., Bai, W., Gao, Y., Dong, X., & Lv, W. (2021). Multi-Focus Image Fusion Based on Convolution Neural Network for Parkinson's Disease Image Classification. *Diagnostics*, *11*(12), 2379.

[21] Ruby, R., Zhong, S., ElHalawany, B. M., Luo, H., & Wu, K. (2021). SDN-enabled energy-aware routing in underwater multi-modal communication networks. *IEEE/ACM Transactions on Networking*, *29*(3), 965-978.

[22] Wu, C., Li, X., Guo, Y., Wang, J., Ren, Z., Wang, M., & Yang, Z. (2022). Natural language processing for smart construction: Current status and future directions. *Automation in Construction*, *134*, 104059.

[23] Teo, K. Y., Daescu, O., Cederberg, K., Sengupta, A., & Leavey, P. J. (2022). Correlation of histopathology and multi-modal magnetic resonance imaging in childhood osteosarcoma: Predicting tumor response to chemotherapy. *Plos one*, *17*(2), e0259564.

[24] Švec, J., Neduchal, P., & Hrúz, M. (2022). Multi-modal communication system for mobile robot. *IFAC-PapersOnLine*, *55*(4), 133-138.

[25] Hu, S. (2022). Analysis and Research on Oil Production in Ultra High Water Cut Stage Based on Iots. In *International Conference on Multi-modal Information Analytics* (pp. 1005-1010). Springer, Cham.

[26] Dhabliya, D. Delay-Tolerant Sensor Network (DTN) Implementation in Cloud Computing (2021) Journal of Physics: Conference Series, 1979 (1), art. no. 012031,

[27] Agrawal, S.A., Umbarkar, A.M., Sherie, N.P., Dharme, A.M., Dhabliya, D. Statistical study of mechanical properties for corn fiber with reinforced of polypropylene fiber matrix Composite (2021) Materials Today: Proceedings,