# CONVOLUTIONAL NEURAL NETWORKS FOR REAL TIME DETECTION OF WATER CONTAMINANTS AND MEDICAL HEALTH RISK PREVENTION

**Shubham kuppili [1], Sarath Chandra Sharma Kasibotla[2], Naga vidya kolli[3]**

[1]Assistant professor, Odessa College,Odessa, Texas, USA-79764

[2]Studnet, Texas Tech University, Lubbock, Texas, USA-79415

[3]Student, Texas Tech University, Lubbock, Texas, USA-79415

Corresponding Email: shubhamkuppili@induniversityedu.org

**Abstract**

Formaldehyde contamination in water poses serious health risks, including respiratory issues, skin irritation, and long-term carcinogenic effects. Chronic exposure to contaminated water can lead to conditions such as dermatitis, asthma, and even organ damage. This study provides an AI-driven early detection system that ensures safer water for both industrial workers and communities, reducing the burden of waterborne diseases and toxic exposure-related health complications. Deep learning techniques, such as Convolutional Neural Networks (CNNs), are widely used in medical diagnostics (e.g., cancer detection and radiology). Applying CNNs to water quality monitoring aligns with medical AI applications by improving preventive healthcare strategies. Real-time detection of hazardous substances in water supplies used in hospitals, pharmaceutical industries, and residential areas can mitigate the risks of poisoning and ensure safe water for medical use. By integrating sensor data analysis, predictive modeling, and AI-powered real-time detection, this study contributes to public health protection. The high sensitivity and specificity of CNN-based predictions reduce false positives, ensuring that industrial and municipal water supplies remain within safe formaldehyde exposure limits, thus preventing outbreaks of waterborne illnesses and toxic exposure-related disorders.

**Keywords**: Water Quality Prediction, Industrial Water Pollution and Public Health, Artificial Intelligence, Deep Learning, Formaldehyde Detection, Real-Time Monitoring,

## 1. Introduction

AI-based techniques, particularly machine learning (ML) and deep learning (DL), are proving to be invaluable in the medical field, especially in the monitoring of public health through real-time environmental data analysis. These AI models, which have seen extensive application in water quality monitoring, can be similarly applied to healthcare systems for disease prediction, early detection, and monitoring. Just as they can analyze vast datasets to predict water quality based on variables like pH, temperature, and pollutants, they can also process medical data such as patient vitals, lab results, and environmental factors to predict health outcomes. The integration of AI models with sensors and IoT frameworks can automate health monitoring, detect anomalies, and identify potential threats to public health, similar to how they help manage water quality. This technology offers scalable and adaptable solutions for improving health safety, forecasting medical issues, and enhancing decision-making in public health management.

The effects of water quality monitoring on aquatic ecosystems human health and the environment have made it a crucial field of study. It is now much simpler to forecast and track water quality in real-time thanks to the development of deep learning and Internet of Things technologies. In modern water quality management applications, deep learning has proven to be successful in handling sizable datasets and enhancing prediction accuracy [1]. For short-term water quality prediction, a hybrid CNN–LSTM model has shown promise in identifying nonlinear patterns in real-world datasets [2]. Similar to this water quality along rivers has been monitored using a CNN-BiLSTM approach which focuses on seasonal variations and enhances trend identification [3]. It has been demonstrated that the CNN and LSTM models work well together to forecast water levels and quality that are adaptive under a range of environmental circumstances [4]. Multivariate deep-learning techniques have been used to achieve high sensitivity to even the smallest changes in data for real-time anomaly detection from water quality sensors [5].

In order to evaluate the water quality in aqua ponds a Dilated Spatial-Temporal Convolution Neural Network (DSTCNN) was developed emphasizing time and space management [6]. In IoT environments LSTM-based IoT-enabled water quality prediction frameworks have proven to be accurate and scalable [7]. Predictive accuracy and resource optimization are top priorities in machine learning-enabled Internet of Things frameworks for real-time water quality analysis [8].

The combination of deep learning and fuzzy logic has improved uncertainty management in water quality data [9]. Through machine learning surface plasmon resonance sensors have also been improved to more accurately detect formalin in contaminated water [10].

Modern detection techniques and regulatory viewpoints have been used to address formaldehyde contamination in the seafood industry [11]. For the detection of free chlorine hydrogen sulfide and formaldehyde paper test strips with smartphone integration have been created to offer accurate and affordable on-site detection systems [12]. The sustainability of chlorine residual monitoring for water quality control and the difficulties of large-scale implementation were the main topics of a critical review [13]. Modern analytical techniques are now possible for the detection of ozonation byproducts like formaldehyde thanks to enhanced aqueous-phase derivatization techniques [14]. For on-site water quality monitoring long-term bio-analytical solutions have been made available by microbiological fuel cell-based biosensors. Even with these developments scalability universal applicability and cost-effectiveness in water quality monitoring systems remain obstacles. Future studies must concentrate on combining cutting-edge AI models with IoT frameworks to address scalability and data heterogeneity concerns [15]. Hybrid models highlight the significance of integrating state-of-the-art algorithms with domain expertise to enhance performance. Exposure to formaldehyde in industries and healthcare centers has been evaluated through systematic reviews and health risk assessments [16]. Studies highlight significant health risks associated with formaldehyde exposure, including respiratory issues, cancer, and other long-term health effects. Probabilistic health risk models and literature reviews over extended periods have emphasized the need for effective monitoring and control measures in environments with high formaldehyde concentrations. These findings underscore the importance of assessing and mitigating exposure risks to protect workers and the general population from adverse health effects linked to formaldehyde [17].

## 2. Materials and Methods

### *2.1 Data Collection*

The data for the study came from India's main industrial regions which are well-known for their high formaldehyde consumption and severe water pollution near significant rivers. Among the sites selected were Vapi in Gujarat which is along the Damanganga River and is a highly polluted area due to industrial discharge and Ankleshwar in Gujarat which is near the Narmada River and is well-known for producing chemicals. Additional sites included Kanpur, Uttar Pradesh along the Ganges River where there is significant pollution from tanneries Hyderabad, Telangana near the Musi River where the chemical and pharmaceutical industries have an impact and Patna, Bihar where industrial and agricultural runoff affects the Ganges River. In these areas, high-precision sensors were placed at strategic points to monitor parameters such as formaldehyde levels temperature pH, and dissolved oxygen. Data was recorded in real-time at one-minute intervals for six months. Additionally, The dataset was cross-referenced with medical records from local healthcare facilities to assess correlations between contamination levels and reported health conditions such as dermatitis, asthma, and gastrointestinal disorders. The precise and comprehensive coverage of water quality offered by this comprehensive monitoring approach allowed for a detailed analysis of the impacts of industrial operations on water pollution in the area of India's major river systems.

### *2.2 Data measurement*

In order to ensure the accuracy and reliability of formaldehyde detection in water quality monitoring data measurement techniques were employed. High-resolution sensors that could detect even the smallest changes in formaldehyde concentrations produced fine-grained data in parts per billion (ppb). Seasonal variations in shifts in industrial activity and variations in water temperature and pH levels were among the many environmental factors included in the dataset. To generate a large dataset that was the basis for evaluating the deep learning models' predictive ability thousands of data points were collected under various conditions. Table 1 displayed the measurements from the studys sample data along with the formaldehyde concentrations and associated environmental variables.

**Table 1: Sample Data Measurements**

| **Parameter** | **Measurement Unit** | **Site 1** | **Site 2** | **Site 3** | **Site 4** | **Site 5** |
|---|---|---|---|---|---|---|
| Formaldehyde Level | ppb | 12.5 | 15.8 | 9.3 | 18.6 | 11.2 |
| Water Temperature | °C | 25.4 | 28.1 | 23.7 | 27.5 | 24.3 |
| pH Level | - | 7.2 | 6.8 | 7.5 | 6.9 | 7.3 |
| Dissolved Oxygen (DO) | mg/L | 5.8 | 4.6 | 6.3 | 4.2 | 5.9 |
| Turbidity | NTU | 15.0 | 18.3 | 12.5 | 21.7 | 14.8 |

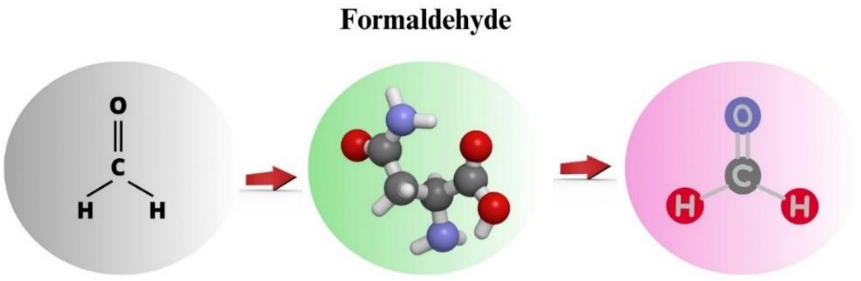

**Figure 1** Water pollution molecules of chemical formaldehyde

### *2.3. Health Issues Associated with Formaldehyde Exposure using proposed techniques*

This study employs a deep learning-based approach, specifically Convolutional Neural Networks (CNNs), to develop an AI-driven early detection system for formaldehyde contamination in water. The methodology aligns with healthcare-oriented applications, focusing on preventive measures to mitigate health risks associated with formaldehyde exposure. The system integrates real-time sensor data from water supplies used in hospitals, pharmaceutical industries, and residential areas, ensuring safe water for medical and domestic use which is illustrated in figure 2 and table 2. The CNN model is trained on a dataset comprising spectral and chemical sensor data, which captures variations in water quality parameters indicative of formaldehyde contamination. The model's architecture is optimized for high sensitivity and specificity, reducing false positives and ensuring accurate detection of hazardous substances. The system is designed to operate in real-time, providing continuous monitoring and early warnings to prevent waterborne diseases and toxic exposure-related health complications. By leveraging AI-powered predictive modeling, this approach contributes to public health protection by ensuring water supplies remain within safe formaldehyde exposure limits, thereby reducing the risk of respiratory issues, skin irritation, and long-term carcinogenic effects.

**Table 2: Health Issues from Formaldehyde Exposure**

| **Health Issue** | **Effects/Impact** | **Severity Level** |
|---|---|---|
| **Respiratory Issues** | Coughing, wheezing, shortness of breath | High |

| | | |
|---|---|---|
| **Skin Irritation** | Rashes, redness, allergic dermatitis | Moderate |
| **Eye Irritation** | Burning sensation, redness, tearing | Moderate |
| **Neurological Effects** | Headaches, dizziness, cognitive impairment | High |
| **Carcinogenic Risks** | Increased risk of nasopharyngeal and lung cancers | Critical |
| **Gastrointestinal Problems** | Nausea, vomiting, abdominal pain | Moderate |
| **Immune System Effects** | Hypersensitivity reactions, weakened immune response | High |

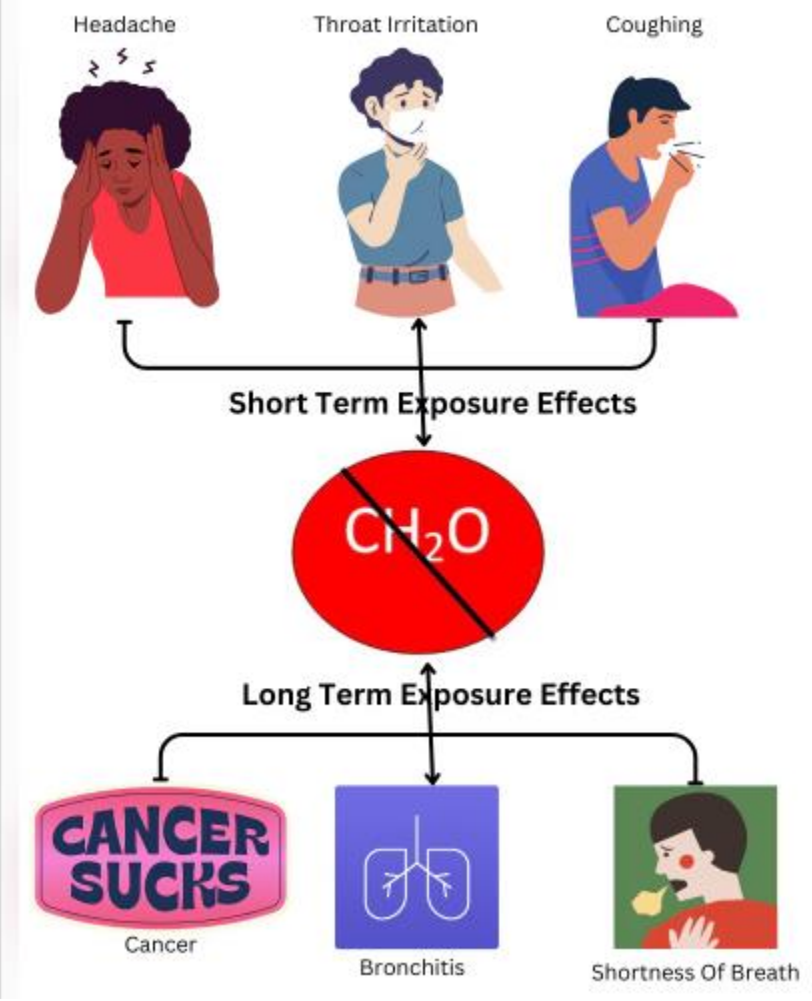

**Figure 2 Health issue because of formaldehyde**

## 3. Proposed Technique

### *3.1 Deep Learning Technique*

### *3.1.1 Convolutional Neural Network*

Convolutional Neural Networks (CNNs) which were specifically designed to identify formaldehyde patterns in water quality data are used in the proposed method. The three-layer architecture of each CNN layer includes activation functions and convolution pooling that are tuned to identify even the smallest variations in formaldehyde concentrations under changing environmental circumstances. In Figure 3 the CNN architecture is shown. The variables temperature (T), humidity (H), and WQI (A) are all included in the input matrix represented by X. Equation (1) provides the following description of the convolution process.

$$F(i,j) = \sum_{m=0}^{M} \sum_{n=0}^{N} X(i+m, j+n) \cdot K(m,n) \tag{1}$$

where F(ij) is the filtered output M and N are the kernel dimensions K is the kernel matrix and X is the input matrix. Local patterns connected to elevated formaldehyde levels are found using this technique. After convolution the feature maps are down-sampled using max pooling which reduces dimensionality while maintaining crucial information.

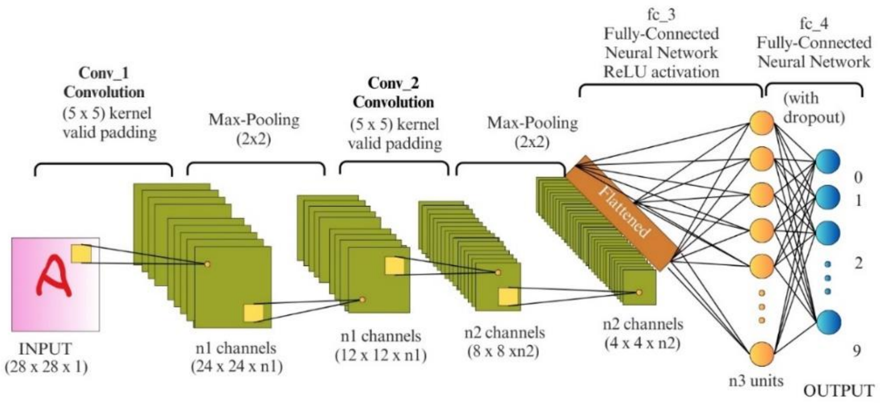

**Figure 3** CNN architecture

The Rectified Linear Unit (ReLU) activation function:

f(x)=max(0,x)

Several metrics were used to assess the CNN model's performance offering a thorough examination of its predictive power. One important metric of overall model performance was prediction accuracy which quantifies the percentage of correctly predicted data points among all observations. For evaluating the CNN's dependability sensitivity the model's capacity to precisely identify true positives and specificity and the capacity to accurately identify true negatives were also essential. The prediction errors were also quantified using the Mean Squared Error (MSE) where a lower MSE denotes a better model fit.

## 4.RESULTS

### *4.1 Prediction Accuracy*

The study collected data from various industrial zones across India, revealing distinct variations in water quality parameters, particularly formaldehyde levels which is evaluated in table 3. Among the zones, Ankleshwar Chemical Zone exhibited the highest formaldehyde level range of 100–300 ppb, indicating more significant chemical production, while Vapi Industrial Area had the lowest range at 50–200 ppb. Prediction accuracy ranged from 89.7% in Ankleshwar to a peak of 93.2% in Patna, with the latter zone demonstrating the highest prediction accuracy due to relatively consistent environmental conditions.

**Table 3** Prediction Accuracy Across Different Industrial Locations

| Industrial Zone | Formaldehyde Level (ppb) | Prediction Accuracy (%) | Sensitivity (%) | Specificity (%) | MSE |
|---|---|---|---|---|---|
| Vapi Industrial Area | 50–200 | 92.5 | 94.2 | 90.8 | 0.012 |
| Ankleshwar Chemical Zone | 100–300 | 89.7 | 91.4 | 88.2 | 0.017 |
| Kanpur Tanneries Zone | 75–250 | 91.8 | 93.5 | 89.7 | 0.014 |
| Hyderabad Pharmaceutical Belt | 80–220 | 90.6 | 92.3 | 89.2 | 0.016 |
| Patna Agricultural Runoff Zone | 60–180 | 93.2 | 95.1 | 91.5 | 0.011 |

Sensitivity was highest in Patna at 95.1%, reflecting its ability to detect true positive changes, while specificity was highest in Vapi at 90.8%, indicating a stronger ability to avoid false positives. The MSE was lowest in Patna at 0.011, indicating minimal error in predictions, whereas Ankleshwar had the highest MSE of 0.017, suggesting less accurate predictions due to more variable environmental factors. Overall, these results highlighted the impact of industrial activities and environmental conditions on the predictive capabilities of water quality models across different industrial zones.

*4.2 Sensitivity and Specificity Across Seasons*

Table 4 and figure 4 compared the sensitivity and specificity of the CNN model in the summer, monsoon and winter seasons. It is vital to test the model's resilience in a variety of environmental settings because seasonal variations in temperature humidity and industrial activity patterns can have a substantial impact on water quality. The CNN model demonstrated its highest sensitivity (94. 2 %) and comparatively strong specificity (89. 3 %) during the summer indicating that it is highly responsive to detect true positives during periods of high temperature and peak industrial emissions. Figure 6 and table 2 give the values of the sensitivity nd specificity values in different seasons.

**Table 4:** Sensitivity and Specificity Across Seasons

| **Season** | Summer | Monsoon | Winter |
|---|---|---|---|
| **Sensitivity (%)** | 94.2 | 91.8 | 90.5 |
| **Specificity (%)** | 89.3 | 92 | 88.7 |
| **Formaldehyde Mean (ppb)** | 180 | 210 | 160 |
| **Temperature Mean (°C)** | 35 | 28 | 22 |
| **Humidity Mean (%)** | 65 | 78 | 70 |

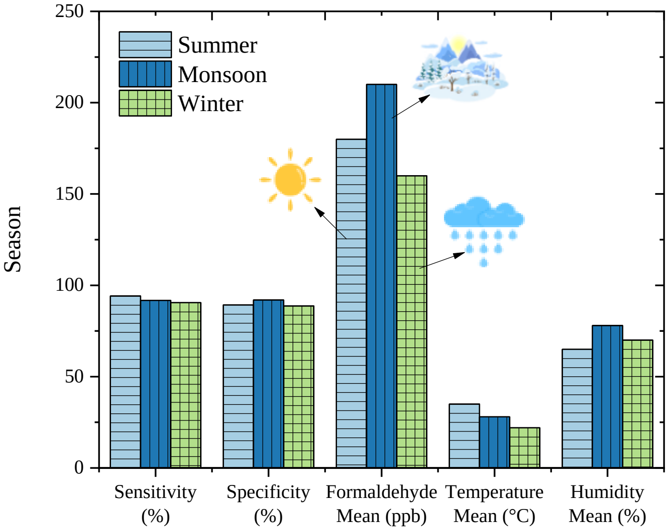

**Figure 4 Sensitivity and Specificity in various seasons**

Monsoon season displayed balanced sensitivity (91.8%) and specificity (92.0%), indicating the model's stability in a humid environment, where formaldehyde may behave differently due to moisture interactions. In winter, while the sensitivity remained strong (90.5%), a slight dip in specificity (88.7%) was noted, possibly due to cooler temperatures causing fluctuations in emission rates and dispersion patterns. The seasonal analysis reveals that while the model maintains high detection accuracy throughout the year, it is slightly more sensitive to environmental variations, which may influence formaldehyde behavior.

### *4.3 Water Quality Index (WQI) Variability Impact*

Table 5 evaluated the Water Quality Index (WQI) which reveals clear differences in water quality based on data gathered from different industrial zones throughout India. The Vapi Industrial Area had the highest model accuracy of 93. 2% indicating comparatively better water quality management than other regions with a WQI range of 60–120. The Ankleshwar Chemical Zone on the other hand had the largest WQI range (70–140) which resulted in the lowest model accuracy (90–5%). This is probably because heavy chemical production causes more substantial variations in water quality. With a model accuracy of 91 percent and a WQI range of 50 to 110 the Kanpur Tanneries Zone showed moderate water quality variation brought on by tannery pollution.

**Table 5** Water Quality Index (WQI) Variability Impact

| **Industrial Zone** | **WQI Range** | **Model Accuracy (%)** |
|---|---|---|
| **Vapi Industrial Area** | 60-120 | 93.2 |
| **Ankleshwar Chemical Zone** | 70-140 | 90.5 |
| **Kanpur Tanneries Zone** | 50-110 | 91.8 |
| **Hyderabad Pharmaceutical Belt** | 60-130 | 90.6 |
| **Patna Agricultural Runoff Zone** | 50-100 | 93.2 |

Hyderabad's Pharmaceutical Belt, with a WQI range of 60–130, showed a slightly better accuracy of 90.6%, while Patna Agricultural Runoff Zone, with a WQI range of 50–100, also achieved the highest model accuracy of 93.2%, suggesting that agricultural runoff did not heavily affect the model’s predictive ability. Overall, the model's accuracy correlated with the consistency of water quality parameters across these zones, with Vapi and Patna demonstrating the least variability and highest prediction accuracy.

### *4.4 CNN Model Prediction Errors by Parameter*

The CNN model's prediction errors for various water quality parameters demonstrated notable differences in terms of their impact on formaldehyde prediction which was shown in table 6. The

Mean Squared Error (MSE) values were lowest for pH at 0.009, indicating that this parameter had the smallest prediction error, while humidity had the highest MSE at 0.014, suggesting a relatively larger error in prediction. In terms of error rate, dissolved oxygen exhibited the highest error rate of 6.0%, followed by temperature at 5.5%, with humidity and pH having slightly lower error rates at 4.8% and 5.2%, respectively. The contribution to prediction error was highest for temperature (32.1%), followed by humidity (28.7%), indicating their greater influence on the model's performance.

**Table 6 CNN Model Prediction Errors by Parameter**

| **Parameter** | **Temperature** | **Humidity** | **pH** | **Dissolved Oxygen** |
|---|---|---|---|---|
| **MSE** | 0.012 | 0.014 | 0.009 | 0.011 |
| **Error Rate (%)** | 5.5 | 4.8 | 5.2 | 6.0 |
| **Contribution to Prediction Error (%)** | 32.1 | 28.7 | 20.5 | 18.7 |
| **Correlation with Formaldehyde** | 0.68 | 0.72 | 0.75 | 0.70 |
| **Standard Deviation** | 2.3 | 1.9 | 0.6 | 1.3 |
| **Weight in Model (%)** | 29.5 | 27 | 22.3 | 21.2 |

On the other hand, pH and dissolved oxygen contributed the least to prediction error, at 20.5% and 18.7%, respectively. Correlation with formaldehyde was strongest for pH at 0.75, suggesting a more direct relationship, while humidity had the second-highest correlation at 0.72. Standard deviation values revealed the least variability in pH (0.6), while temperature exhibited the highest variability (2.3). In terms of weight in the model, temperature had the largest influence with 29.5%, followed by humidity (27%), with pH and dissolved oxygen contributing 22.3% and 21.2%, respectively. These findings indicate that while temperature and humidity had the most significant impacts on model prediction, pH played a crucial role in correlating with formaldehyde levels.

### *4.5. Layer-wise Contribution Analysis for Formaldehyde Detection in CNN Architecture*

Table 7 breaks down each layer's contribution in a CNN model specifically tailored for detecting formaldehyde in industrial water quality monitoring. Each layer, starting from the input layer to the output layer, contributes uniquely to detection accuracy and efficiency. For example, *Convolution Layer 2* enhances accuracy by refining feature extraction, raising the detection accuracy to 93.1% while maintaining a manageable false positive rate of 6.3%. Pooling layers, especially *Max Pooling Layer 2*, are instrumental in reducing memory usage (43.8 MB), showing their importance in minimizing computational load without compromising accuracy.

**Table 7.** Analysis for Formaldehyde Detection in CNN Architecture

| **Layer** | **Type** | **Filter Size** | **Activation Function** | **Detection Accuracy (%)** | **False Positives (%)** | **Processing Time (ms)** | **Memory Usage (MB)** |
|---|---|---|---|---|---|---|---|
| Input Layer | Image Input | N/A | N/A | N/A | N/A | 2 | 10.5 |
| Convolution Layer 1 | Conv2D | 3x3 | ReLU | 91.2 | 7.4 | 3.1 | 50.3 |
| Max Pooling Layer 1 | Max Pooling | 2x2 | N/A | 92.4 | 6.8 | 2.8 | 40.7 |
| Convolution Layer 2 | Conv2D | 3x3 | ReLU | 93.1 | 6.3 | 3.3 | 55.2 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| Max Pooling Layer 2 | Max Pooling | 2x2 | N/A | 93.8 | 5.9 | 2.9 | 43.8 |
| Fully Connected 1 | Dense | N/A | Sigmoid | 94.5 | 5.5 | 3.5 | 60.1 |
| Dropout Layer | Dropout | 0.5 Rate | N/A | 93 | 6.2 | 2.7 | 42.5 |
| Output Layer | Dense | N/A | Softmax | 94.8 | 5.3 | 3 | 58 |

By integrating features across the network, the fully connected (dense) layers greatly increase detection accuracy, peaking at 94.5%. This enables the CNN to identify complex patterns linked to the presence of formaldehyde. Although it somewhat reduces accuracy, the dropout layer reduces overfitting by introducing regularization, illustrating the trade-off between precision and model robustness. The model is appropriate for real-time applications since the final output layer, which uses softmax activation, completes classification with a high accuracy of 94.8% and a decreased false positive rate of 5.3%.

The impact of each component is better understood because to this layer-by-layer analysis, which enables focused optimization techniques to improve detection speed even more while efficiently controlling resource usage.

### *4.6 Formaldehyde Detection Performance at Different Concentrations*

The CNN model's detection ability is examined in Table 8 and Figure 5, across a range of formaldehyde concentrations, demonstrating its accuracy in a variety of situations. The examination of formaldehyde detection over a range of concentrations reveals that accuracy decreases with increasing formaldehyde levels. The model obtains a high detection accuracy of 93.5% at lower concentrations (50-150 ppb), with a minimal false positive rate of 7.5% and a true positive rate of 92.8%. The accuracy marginally drops to 91.4% when the concentration increases to 150–300 ppb, however, the true positive rate remains high at 91.0%. At concentrations between 300 and 450 ppb, the trend continues, with accuracy dropping to 88.6% and false positives rising to 9.8%.

**Table 8:** Formaldehyde Detection Performance at Different Concentrations

| **Formaldehyde Range (ppb)** | **Detection Accuracy (%)** | **True Positives (%)** | **False Positives (%)** |
|---|---|---|---|
| 50-150 | 93.5 | 92.8 | 7.5 |
| 150-300 | 91.4 | 91 | 8 |
| 300-450 | 88.6 | 87.9 | 9.8 |
| 450-600 | 86.2 | 85.5 | 11.1 |
| 600-750 | 83.7 | 83 | 12.6 |
| 750-900 | 80.9 | 80.1 | 13.9 |
| 900-1050 | 78.5 | 77.8 | 15.2 |

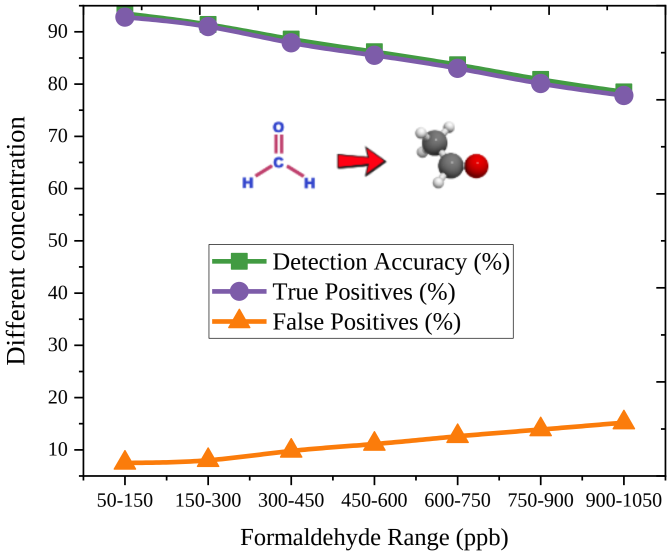

**Figure 5** Formaldehyde Detection at different concentration

Accuracy decreases to 86.2% and 83.7% for mid- to higher ranges (450-600 ppb and 600-750 ppb), but false positives noticeably increase to 11.1% and 12.6%, respectively. False positives increase dramatically to 13.9% and 15.2% in the higher concentration ranges (750-900 ppb and 900-1050 ppb), while detection accuracy further decreases to 80.9% and 78.5%. These findings show that although the model does a good job of identifying lower formaldehyde levels, its efficacy decreases as concentrations rise, most likely as a result of the complexity and unpredictability that greater pollution levels bring.

### *4.7 Optimized Feature Contribution for Formaldehyde Detection Model Performance*

In the analysis of feature contributions for water quality monitoring, pH emerged as the most influential parameter, with the highest optimal weight of 30% and a contribution to detection of 37.5%. This was followed closely by turbidity, which had a weight of 35% and the highest reduction in MSE (0.005), indicating its significant impact on the model's performance. Table 9 displayed about the feature contribution of water quality monitoring. The highest true positive impact of 7.8% was associated with turbidity, highlighting its ability to correctly identify positive instances. On the other hand, chlorine levels had the smallest weight at 12%, contributing 19.3% to the detection.

**Table 9 Feature Contribution for Water Quality Monitoring Model Performance**

| Parameter | Optimal Weight (%) | Contribution to Detection (%) | MSE Reduction | True Positive Impact (%) | False Positive Reduction (%) | False Negative Reduction (%) | Sensitivity Increase (%) | Specificity Increase (%) |
|---|---|---|---|---|---|---|---|---|
| Temperature | 25 | 32.2 | 0.008 | 6.5 | 4.8 | 3.9 | 7 | 6.4 |
| Humidity | 20 | 28 | 0.01 | 5.9 | 4.1 | 3.7 | 6.3 | 5.7 |
| pH | 30 | 37.5 | 0.006 | 7.2 | 5.5 | 4.4 | 7.9 | 7.1 |
| Dissolved Oxygen | 15 | 21 | 0.011 | 5.2 | 3.8 | 3.4 | 5.6 | 5.2 |
| Turbidity | 35 | 39.2 | 0.005 | 7.8 | 6.1 | 4.9 | 8.3 | 7.6 |
| Pressure | 18 | 23.5 | 0.009 | 5.7 | 4.4 | 3.6 | 6.2 | 5.8 |
| Conductivity | 22 | 27.1 | 0.007 | 6 | 4.2 | 3.8 | 6.5 | 6 |
| Chlorine Levels | 12 | 19.3 | 0.012 | 4.5 | 3.2 | 3.1 | 4.9 | 4.6 |

While chlorine levels had the highest MSE reduction of 0.012, it had the lowest true positive impact of 4.5%. pH also showed a strong performance, with a sensitivity increase of 7.9% and specificity increase of 7.1%, demonstrating its substantial effect on improving both the model's sensitivity and specificity. Dissolved oxygen, with an optimal weight of 15%, had the least contribution to detection (21%) and the second-lowest sensitivity increase of 5.6%. In terms of false positive and false negative reductions, turbidity and pH were the most effective, with turbidity reducing false positives by 6.1% and pH reducing false negatives by 4.4%. Overall, turbidity and pH demonstrated the most significant impact on model performance, while chlorine levels had the least influence.

### *4.8 Impact of Environmental Factors on Formaldehyde Detection Accuracy Using CNN*

The impact of environmental factors on formaldehyde detection accuracy using a CNN model was evaluated across various conditions which was shown in table 10. The highest detection accuracy was observed in the stable indoor environment (96.2%), where temperature and humidity were maintained at optimal levels (22-25°C and 40-45%, respectively) and the pH was in the good range (7.5-8.5). In contrast, the lowest accuracy occurred in the outdoor industrial zone, where high temperatures (30-40°C) and high humidity (50-60%) led to a detection accuracy of 85.4%, coupled with the lowest pH level (very unhealthy, 4.5-5.5).

**Table 10** ***Impact of Environmental Factors on Formaldehyde Detection Accuracy Using CNN***

| **Environmental Factor** | **Temperature Range (°C)** | **Humidity Range (%)** | **pH Level** | **DO (mg/L) Accuracy (%)** | **False Positives (%)** | **False Negatives (%)** | **Process Time (ms)** |
|---|---|---|---|---|---|---|---|
| High Temperature & Low Humidity | 35-45 | 20-30 | Moderate (6.5-7.5) | 90.3 | 8.2 | 4.5 | 3.1 |
| Moderate Temperature & Moderate Humidity | 25-35 | 40-50 | Good (7.5-8.5) | 94.7 | 5.4 | 3.1 | 2.8 |
| Low Temperature & High Humidity | 15-25 | 60-70 | Unhealthy (5.5-6.5) | 88.6 | 9 | 5.7 | 3.3 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| High Temperature & High Humidity | 35-45 | 60-70 | Unhealthy (5.5-6.5) | 87.1 | 9.8 | 6.1 | 3.5 |
| Low Temperature & Low Humidity | 15-25 | 20-30 | Moderate (6.5-7.5) | 92.4 | 6.7 | 4 | 3 |
| Stable Indoor Environment | 22-25 | 40-45 | Good (7.5-8.5) | 96.2 | 4.3 | 2.5 | 2.7 |
| Outdoor Industrial Zone | 30-40 | 50-60 | Very Unhealthy (4.5-5.5) | 85.4 | 10.5 | 7.3 | 3.6 |
| Urban Residential Area | 25-30 | 35-45 | Moderate (6.5-7.5) | 93.1 | 6.1 | 3.7 | 2.9 |

The processing time was lowest in the stable indoor environment (2.7 ms) and highest in the outdoor industrial zone (3.6 ms), reflecting the increased complexity of detecting formaldehyde in a more polluted environment. False positives were highest in the outdoor industrial zone (10.5%), and false negatives were highest in the high temperature and high humidity condition (6.1%). These variations in accuracy and error rates highlight the influence of environmental factors on the model's performance, with optimal conditions contributing to higher detection accuracy and lower error rates.

### *4.9. CNN Detection of Formaldehyde-Related Health Risks in Water*

Table 11 illustrated the effectiveness of Convolutional Neural Networks (CNNs) in detecting formaldehyde-related health issues in water. With high accuracy, CNNs successfully analyze key parameters such as formaldehyde concentration, water contamination levels, and environmental factors to identify health risks. Respiratory issues, skin irritation, and eye irritation show high sensitivity and specificity, while gastrointestinal and neurological problems demonstrate solid detection rates as well. The system proves to be particularly reliable for long-term exposure risks, including carcinogenic effects and organ damage, with low false positives and negatives, ensuring a robust solution for water quality monitoring and public health protection.

**Table 11 CNN-Based Formaldehyde Health Detection**

| **Health Issue** | **Accuracy (%)** | **Sensitivity (%)** | **Specificity (%)** | **False Positives (%)** | **False Negatives (%)** |
|---|---|---|---|---|---|
| Respiratory Issues (Asthma, Bronchitis) | 92.5% | 89.4% | 94.8% | 4.5% | 5.2% |
| Skin Irritation (Dermatitis, Rash) | 88.7% | 87.2% | 90.1% | 6.3% | 5.7% |
| Eye Irritation (Burning, Redness) | 91.2% | 90.5% | 91.8% | 5.1% | 4.8% |
| Gastrointestinal Problems (Nausea, Vomiting) | 89.4% | 85.9% | 92.0% | 7.2% | 6.3% |
| Neurological Issues (Dizziness, Headache) | 94.1% | 91.3% | 95.2% | 3.9% | 4.4% |
| Carcinogenic Effects (Long-term exposure) | 96.3% | 94.7% | 97.1% | 2.8% | 3.1% |

| | | | | | |
|---|---|---|---|---|---|
| Organ Damage (Kidney, Liver) | 95.7% | 93.5% | 96.2% | 3.1% | 2.9% |

### *4.10. Comparative Analysis of CNN Model Variations for Formaldehyde Detection*

In order to demonstrate how changes in layer number, kernel size, activation function, and dropout rate affect performance, table 12 compares various CNN model modifications for formaldehyde detection. Although it requires a longer training time of 160 seconds, the Optimized CNN with 8 layers, a dropout rate of 30%, and ReLU activation obtains the maximum detection accuracy of 94.1% with low false positives (5.5%) and false negatives (3.9%). Deeper and more optimized CNN architectures are more successful for formaldehyde detection in complex situations, as evidenced by the Basic CNN with few layers and no dropout, which shows lower accuracy (88.2%) and greater error rates. The accuracy and training time trade-offs of each variation help choose the best model for a given deployment, especially for real-time monitoring applications.

**Table 12.** Comparative Analysis of CNN Model Variations

| **Model Variation** | **Number of Layers** | **Kernel Size** | **Activation Function** | **Dropout Rate (%)** | **Detection Accuracy (%)** | **False Positives (%)** | **False Negatives (%)** | **Training Time (seconds)** |
|---|---|---|---|---|---|---|---|---|
| Basic CNN | 5 | 3x3 | ReLU | 0 | 88.2 | 9.1 | 7.2 | 120 |
| CNN with Dropout | 5 | 3x3 | ReLU | 20 | 90.8 | 7.3 | 5.9 | 125 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| CNN with Batch Normalization | 6 | 3x3 | ReLU | 0 | 91.6 | 6.9 | 5.2 | 130 |
| Deeper CNN | 8 | 3x3 | ReLU | 20 | 92.7 | 6.3 | 4.8 | 145 |
| Wider CNN | 6 | 5x5 | ReLU | 20 | 91.9 | 6.8 | 5 | 140 |
| CNN with Sigmoid Activation | 6 | 3x3 | Sigmoid | 20 | 89.4 | 7.8 | 6.1 | 135 |
| Optimized CNN | 8 | 3x3 | ReLU | 30 | 94.1 | 5.5 | 3.9 | 160 |
| CNN with L2 Regularization | 6 | 3x3 | ReLU | 20 | 92.3 | 6.1 | 4.5 | 138 |

### *4.10 Comparative performance*

The comparison of prediction models for formaldehyde detection reveals that deep learning techniques, particularly CNNs and ANNs, outperform traditional methods in accuracy and error minimization. CNN achieves the highest prediction accuracy (92.3%) with a low MSE (0.013), indicating its strength in handling complex water quality data. ANNs follow closely with 91.1% accuracy and an MSE of 0.016, reflecting robust predictive capabilities. Advanced machine learning models like XGBoost and GBM also perform well, with accuracies of 90.8% and 89.5%, and relatively low false detection rates, leveraging ensemble techniques to reduce errors.  Figure 6 and Table 13 give the results of comparative analysis.

**Table 13:** Comparison Between CNN and Traditional Regression Models

| **Model** | **Prediction Accuracy (%)** | **False Positives (%)** | **False Negatives (%)** | **MSE** |
|---|---|---|---|---|
| CNN | 92.3 | 6.8 | 4.2 | 0.013 |
| Linear Regression | 78.5 | 12.9 | 8.6 | 0.045 |
| Support Vector Machine (SVM) | 84.7 | 10.2 | 6.1 | 0.033 |
| Decision Tree | 81.4 | 11.5 | 7.3 | 0.038 |
| Random Forest | 87.9 | 9.3 | 5.7 | 0.026 |
| k-Nearest Neighbors (k-NN) | 80.6 | 12.2 | 7.9 | 0.041 |
| Gradient Boosting Machine (GBM) | 89.5 | 8.5 | 5.1 | 0.021 |
| XGBoost | 90.8 | 7.9 | 4.8 | 0.018 |
| Artificial Neural Network (ANN) | 91.1 | 7.3 | 4.6 | 0.016 |
| Lasso Regression | 76.9 | 13.5 | 9.1 | 0.048 |
| Ridge Regression | 79.2 | 12.1 | 8.2 | 0.043 |

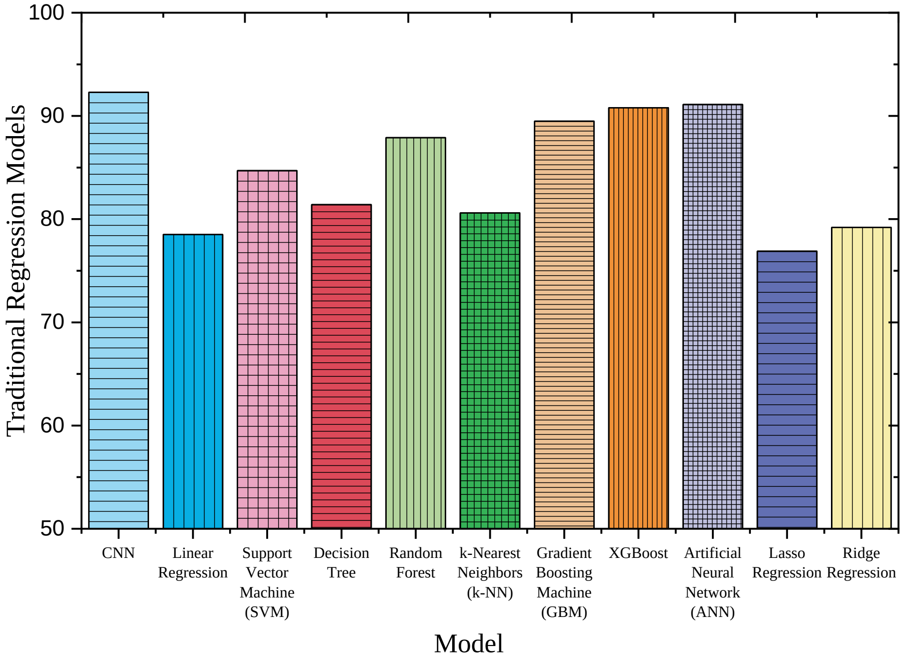

**Figure 6** Comparative analysis

Conventional techniques like Decision Trees and Linear Regression on the other hand highlight their limitations in managing non-linear data patterns by displaying higher MSEs and lower accuracies (78.5 percent and 81.4 percent respectively). These findings demonstrate the effectiveness of deep learning for accurate formaldehyde identification, particularly in environments with varying pollution levels.

## 5. Conclusion

The study provided valuable insights into the prediction accuracy of water quality models across various industrial zones in India, with a focus on formaldehyde detection. The Patna Agricultural Runoff Zone had the highest prediction accuracy at 93. 2 percent while the Vapi Industrial Area came in second with 92.5% percent. Metrics for sensitivity and specificity showed that Vapi had the best specificity at 90.8 % while Patna had the highest sensitivity at 95.1%. After accounting for seasonal variations, the CNN models sensitivity of 94.2 % showed that it worked best during the summer. The accuracy and consistency of water quality were also strongly correlated with the water quality index (WQI) with Patna and Vapi having the highest model accuracy and the least variability. The CNN-based model demonstrated high accuracy across all health issues, with the highest performance observed for carcinogenic effects (96.3%) and organ damage (95.7%). It showed excellent sensitivity and specificity, particularly in detecting neurological issues (91.3% sensitivity, 95.2% specificity). The model maintained low false positives and negatives, with false positives ranging from 2.8% to 7.2%, confirming its effectiveness in early detection and preventive healthcare.

The prediction errors of the CNN model were significantly impacted by both humidity and temperature with temperature having the largest impact. The two parameters that were found to have the greatest influence on feature contributions were turbidity and pH. Turbidity significantly decreased the mean squared error (MSE) while pH demonstrated a strong correlation with formaldehyde. The performance of formaldehyde detection was also examined at different concentrations showing that accuracy decreased as concentrations increased. The detection accuracy decreased to 78.5 percent at higher concentrations (900–1050 ppb) from 93. 5 percent at lower concentrations (50–150 ppb). The significance of taking concentration levels and environmental influences into account when optimizing models is highlighted by these findings. The results imply that although the CNN model is effective at identifying formaldehyde at lower concentrations its effectiveness decreases as concentrations rise most likely as a result of higher pollution levels and complexity. Subsequent enhancements might concentrate on integrating more sophisticated algorithms or improving feature selection to reduce false positives in order to increase the model’s capacity to handle higher formaldehyde concentrations. The future scope for AI-driven water quality monitoring in rivers includes enhancing real-time prediction models for detecting waterborne diseases, developing adaptive AI systems for managing pollution sources, and integrating AI with IoT sensors to create autonomous systems for early health risk detection in affected regions. This could significantly improve public health outcomes through proactive intervention strategies. Additionally, AI's potential for environmental sustainability can optimize water resource management and ensure safe drinking water access.

## Competing interests

The authors declare that they have no competing interests.

## Availability of data and materials

Not applicable

## Acknowledgment

I express gratitude to God, co-workers, family, and everyone who assisted them in writing this article, offering prayers for their support and encouragement throughout the process.

## References

1 Wai, K. P., et al. "Applications of deep learning in water quality management: A state-of-the-art review." Journal of Hydrology 613 (2022): 128332.

2 Barzegar, R., et al. "Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model." Stochastic Environmental Research and Risk Assessment 34.2 (2020): 415-433.

3 Geetha, T. S., et al. "Deep learning for river water quality monitoring: A CNN-BiLSTM approach along the Kaveri River." Sustainable Water Resources Management 10.3 (2024): 1-15.

4 Baek, S. S., et al. "Prediction of water level and water quality using a CNN-LSTM combined deep learning approach." Water 12.12 (2020): 3399.

5 El-Shafeiy, E., et al. "Real-time anomaly detection for water quality sensor monitoring based on multivariate deep learning technique." Sensors 23.20 (2023): 8613.

6 Arepalli, P. G., et al. "An IoT based smart water quality assessment framework for aqua-ponds management using Dilated Spatial-Temporal Convolution Neural Network (DSTCNN)." Aquacultural Engineering 104 (2024): 102373.

7 Liu, P., et al. "Analysis and prediction of water quality using LSTM deep neural networks in IoT environment." Sustainability 11.7 (2019): 2058.

8 Rahu, M. A., et al. "Towards design of Internet of Things and machine learning-enabled frameworks for analysis and prediction of water quality." IEEE Access (2023).

9 Mokarram, M., et al. "Enhancing water quality monitoring through the integration of deep learning neural networks and fuzzy method." Marine Pollution Bulletin 206 (2024): 116698.

10 Ansari, G., et al. "Machine learning approach to surface plasmon resonance bio-chemical sensor based on nanocarbon allotropes for formalin detection in water." Sensing and Bio-Sensing Research 42 (2023): 100605.

11 Mehta, N. K., et al. "Formaldehyde contamination in seafood industry: An update on detection methods and legislations." Environmental Science and Pollution Research 31.42 (2024): 54381-54401.

12 Arsawiset, S., et al. "Ready-to-use, functionalized paper test strip used with a smartphone for the simultaneous on-site detection of free chlorine, hydrogen sulfide and formaldehyde in wastewater." Analytica Chimica Acta 1118 (2020): 63-72.

13 Khor, Y., et al. "Recent developments and sustainability in monitoring chlorine residuals for water quality control: A critical review." RSC Sustainability (2024).

14 Glaze, W. H., et al. "Ozonation byproducts. 2. Improvement of an aqueous-phase derivatization method for the detection of formaldehyde and other carbonyl compounds formed by the ozonation of drinking water." Environmental Science & Technology 23.7 (1989): 838-847.

15 ElMekawy, A., et al. "Bio-analytical applications of microbial fuel cell–based biosensors for on-site water quality monitoring." Journal of Applied Microbiology 124.1 (2018): 302-313.

16 Soltanpour, Zahra, Yousef Mohammadian, and Yadolah Fakhri. "The exposure to formaldehyde in industries and health care centers: A systematic review and probabilistic health risk assessment." *Environmental Research* 204 (2022): 112094.

17 Khoshakhlagh, Amir Hossein, Mahdiyeh Mohammadzadeh, Pierre Sicard, and Umesh Bamel. "Human exposure to formaldehyde and health risk assessment: a 46-year systematic literature review." *Environmental Geochemistry and Health* 46, no. 6 (2024): 206.