applied sciences

MDPI

Article

Detection of Adversarial Attacks against the Hybrid Convolutional Long Short-Term Memory Deep Learning Technique for Healthcare Monitoring Applications

Albatul Albattah [1] and Murad A. Rassam [1,2,*]

[1] Department of Information Technology, College of Computer, Qassim University, Qassim 51452, Saudi Arabia
[2] Faculty of Engineering and Information Technology, Taiz University, Taiz 6803, Yemen
* Correspondence: m.qasem@qu.edu.sa

Abstract: Deep learning (DL) models are frequently employed to extract valuable features from heterogeneous and high-dimensional healthcare data, which are used to keep track of patient well-being via healthcare monitoring systems. Essentially, the training and testing data for such models are collected by huge IoT devices that may contain noise (e.g., incorrect labels, abnormal data, and incomplete information) and may be subject to various types of adversarial attacks. Therefore, to ensure the reliability of the various Internet of Healthcare Things (IoHT) applications, the training and testing data that are required for such DL techniques should be guaranteed to be clean. This paper proposes a hybrid convolutional long short-term memory (ConvLSTM) technique to assure the reliability of IoHT monitoring applications by detecting anomalies and adversarial content in the training data used for developing DL models. Furthermore, countermeasure techniques are suggested to protect the DL models against such adversarial attacks during the training phase. An experimental evaluation using the public PhysioNet dataset demonstrates the ability of the proposed model to detect anomalous readings in the presence of adversarial attacks that were introduced in the training and testing stages. The evaluation results revealed that the model achieved an average F1 score of 97% and an accuracy of 98%, despite the introduction of adversarial attacks.

Keywords: Internet of Healthcare Things (IoHT); anomaly detection; deep learning; convolutional long short-term memory (ConvLSTM); adversarial attacks

check for updates

Citation: Albattah, A.; Rassam, M.A. Detection of Adversarial Attacks against the Hybrid Convolutional Long Short-Term Memory Deep Learning Technique for Healthcare Monitoring Applications. *Appl. Sci.* **2023**, *13*, 6807. https://doi.org/10.3390/app13116807

Academic Editor: Luis Javier Garcia Villalba

Received: 5 May 2023
Revised: 31 May 2023
Accepted: 1 June 2023
Published: 3 June 2023

1. Introduction

The advancement of Internet of Things (IoT) technologies has led to the widespread penetration and large-scale deployment of IoT systems worldwide. While IoT systems are eminently qualified for providing intelligent services, the massive amounts of data collected and processed by IoT systems also raise serious security concerns. Therefore, research efforts have been focused on designing intelligent anomaly detection systems to prevent the misuse of IoT data across smart applications [1].

In the healthcare context, IoT has improved the standard of care accorded to patients. Indeed, people can enjoy life with greater convenience since these systems ensure their health and safety through continuous monitoring. In addition, the IoT supports many healthcare applications, from attached medical sensors to wireless body area networks (WBANs). A WBAN comprises a network of small, wearable devices and is considered the most promising technology for enhancing healthcare services. Such devices have enabled remote monitoring to increase the overall goodness of care provided to patients in remote areas or medical facilities [2,3].

In spite of these advantages, WBANs are also susceptible to external threats since sensor data are gathered from various sources including people and locations. Attackers with malicious motives may target the sensors and insert malicious data that report back anomalous observations, resulting in inaccurate diagnoses, wrong medication being given to patients, and significant financial losses for healthcare entities using this healthcare system [4,5]. Healthcare systems now face a serious problem with anomalies, which can also result from faulty sensors and inaccurate observations from sensing devices.

Anomaly detection is considered one of the best solutions for WBANs in terms of distinguishing abnormal data and may provide a reliable system to counter sensor faults and anomalous activities, along with an understanding of the factors affecting patients and healthcare organizations. Machine learning and statistical techniques have been applied in various studies to detect anomalies over the last few years. Different researchers have studied the employment of these techniques to detect anomalies in WBANs, and their findings support their effectiveness. However, these solutions still have some limitations in providing high accuracy within a limited timeframe without human intervention, in terms of the engineering of features. To overcome this challenge, deep learning is now being used to enhance the performance of anomaly detection models.

Deep learning-based anomaly detection has great potential for increased security, due to its ability to detect anomalous behavior across data sources that cannot be identified using traditional security methods. However, the vulnerability of deep learning models to adversarial attacks is a fundamental obstacle to employing deep learning for security applications [4]. An adversarial attack occurs when adversarial attack data are fed as input into a deep learning model. One example of such an attack is a dataset in which some features have been purposely perturbed to confuse the deep learning model, in order to produce an incorrect prediction [6].

In this regard, the current work investigates the ability of a deep learning model employing a hybrid convolutional long short-term memory (ConvLSTM) approach to identify adversarial attacks in healthcare monitoring systems. The following is a summary of this paper's most significant contributions:

1. Developing a model for detecting anomalies, utilizing the hybrid ConvLSTM technique, to detect adversarial attacks in WBANs. The technique is used to consider an attack in the training and testing phases and shows how adversarial attacks can potentially mislead the anomaly detection model.
2. Evaluating the proposed model when under attack in the training phase, along with both fast gradient sign (FGSM) and basic iterative (BIM) attacks in the testing phase.
3. Developing a proactive method as a countermeasure that will act as a retraining process to strengthen the model itself against adversarial attacks.

The remainder of the paper is organized as follows: Section 2 investigates works in the literature related to WBAN anomaly detection systems and adversarial attacks and includes a discussion of the various points. The suggested model's design and its specific elements are described in Section 3. Section 4 provides an analysis of the findings and the outcomes of the experiment. The authors' conclusions are presented in Section 5, along with some recommendations for further research.

2. Related Works

Increasing emphasis is being placed on anomaly detection in the Internet of Things arena, particularly regarding healthcare systems that create vast amounts of data through WBANs. In the existing studies, numerous models for detecting anomalies in WBAN-based systems using various methods have been suggested, and these will be analyzed in the following paragraphs.

The authors of Ref. [4] assessed the WBAN's dependability in terms of identifying anomalies, assessing human health conditions, hardware failure rates, and transient fault correction mechanisms. The study proposed a measure using the mean time to failure (MTTF), which provided improved performance when assessing the dependability of WBANs in terms of specification and detection of anomalies. The findings indicated a detection rate of 95% and an MTTF of 43.01 s, which is considered to be short. Even though the study indicated a high level of detection as a measure of dependability, the metric must be altered to account for some irregularities.

Another study [5] proposed a methodology for detecting anomalous sensor data. Their approach was built on concurrent fuzzy clustering and data compression using the Hadoop MapReduce model. The experimental results demonstrated that the suggested framework attained better accuracy with the fewest false alarms, achieving between 97% and 98% accuracy. This study employed a parametric statistical method that is computationally intensive.

A different method for recognizing changes in WBAN-collected data was proposed in [7] using the Kalman filter approach. The researchers claimed that this approach can identify any physiological reaction in WBANs. However, the Kalman filter approach has considerable disadvantages, including a greater computational cost to achieve optimal results.

The authors of [8] proposed a unique method for detecting WBAN anomalies, which was based on Gaussian regression and majority voting. Using a real dataset, the proposed method presented a strategy that was capable of differentiating between true medical emergency situations and false alarms. This technique is successful in terms of detection and the false-positive rate, as was demonstrated by the findings. However, this method was hindered by its computing complexity, high false-alarm rate, and sample size. Elsewhere, the authors of [9] proposed enhancing the performance of anomaly detection by using a correlation method for multiple body sensors. The proposed method employed thresholds to detect anomalies. However, in this study, only one kind of correlation that utilized the spatial relationship among sensors was examined, while disregarding the temporal correlation for each sensor reading.

Two additional studies [10,11] examined the unreliability of certain sensors that are responsible for generating many false alarms in medical-based systems. Both investigations used a dynamic sliding window and a weighted moving average to identify abnormally flawed sensor values. However, researchers using the weighted moving average method need to pay more attention to the complex relationships within the data.

The authors of [12] proposed an anomaly detection algorithm for WBANs to eliminate the false alarms resulting from faulty measurements. The strategy utilized both spatial–temporal correlation and a game theory approach. Additionally, the local processing unit of the proposed design utilized the Mahalanobis distance for multivariate analysis. The suggested method demonstrated improved efficacy in producing a reduced false alarm rate and excellent detection precision. One potential flaw of the game-theoretic method is that it is necessary to account for novel anomalies. The experimental findings revealed that the suggested technique offered the quickest execution time and greatest energy efficiency among the sensors. However, this technique is incapable of detecting random changes in various physiological signals.

A study conducted by the authors of [13] evaluated the performance of three intrusion detection models, based on CNN, LSTM, and the gated recurrent unit (GRU), by applying adversarial examples such as the fast gradient sign method (FGSM) to test the robustness of the three models. The experimental outcomes showed that CNN is the model best able to withstand such adversarial examples. The robustness of GRU and LSTM against adversarial examples can be significantly increased after adversarial training.

The authors of [14] suggested an anomaly detection system for a WBAN-based data sampling technique with a modified cumulative sum as the foundation for an anomaly detection system for use with WBAN (known as MCUSUM). The MCUSUM algorithm was used to reliably detect abnormalities, while the sampling technique was used to boost the detection speed. The findings of this study demonstrated that the suggested technique offered the highest energy efficiency and shortest execution time when using the sensors. However, the anomaly detection system was unable to identify random abnormalities in the different sets of physiological data.

In [15], the authors designed a novel adversarial attack for testing DL-based network intrusion detection systems. The study presented two techniques: the first is a model extraction, while the second uses a saliency map. With these techniques, the attack model was successfully compromised, and the malicious packets had an attack success rate of 94.31%, on average.

The research presented in [16] contrasted the deep autoencoder (DAE), the shallow autoencoder (SAE), and the ensemble of autoencoders, three machine learning models that are utilized for anomaly detection. The study assessed the models' resilience in the face of data poisoning attacks (DPAs). The evaluation results showed an F1 score of ≈97% when handling unpoisoned benign traffic. However, when challenged by a DPA, DAE demonstrated more robust detection capabilities, providing over 50% of the F1 score, with 10% poisoning. The other models, however, exhibited a significantly declining performance (down to a 20% F1 Score), with only 0.5% of the malicious traffic being injected into the data.

The authors of [17] proposed a model agnostic explainability-based method for accurately detecting adversarial attacks on two datasets. The study obtained an accuracy of 88% on the MIMIC-CXR dataset, significantly exceeding the state-of-the-art system of adversarial detection in both datasets by over 10% in all scenarios. The study showed a detection accuracy of 77% against a longitudinal adversarial attack.

In [18], the authors proposed a framework designed to detect poisoning attacks, utilizing deep neural networks and support vector machines. The authors evaluated the framework using different state-of-the-art data poisoning attacks for several healthcare applications: human activity recognition and electrocardiograph classification. The experimental analysis shows that the proposed framework could efficiently detect poisoning attacks and remove the specified poisoned updates from global data aggregation.

The authors of [19] presented a new kind of adversarial approach that would take advantage of the ML classifiers employed in a smart healthcare system (SHS). The study manipulated readings from medical devices to alter the patients' status. The test findings showed that the suggested adversarial approach might severely impair the ability of an ML-based SHS to accurately detect the patients' normal activities, which would ultimately result in incorrect therapy. However, machine learning methods may have other disadvantages since the outcome is dependent on the input dataset.

As reported in [20], a cognitive machine learning-assisted attack detection framework (CML-ADF) has been developed to transmit healthcare data securely. The suggested framework achieves a 96.5% attack prediction ratio, a 97.8% efficiency ratio, a 98.2% accuracy ratio, a 21.3% reduction in latency, and an 18.9% reduction in communication cost, according to the experimental data. Elsewhere, the authors of [21] proposed a model that combined the capabilities of hybrid convolutional long short-term memory (ConvLSTM) methods with correlations in the various physiological data intended to detect contextual and datapoint abnormalities in the massive WBAN datastream. The results of the experiments using the proposed model reported an average accuracy of 99% and a 98% F1-measure on various dataset subjects, compared to 64% for the CNN and LSTM when used separately. However, that study did not examine adversarial attacks. As a result, this paper will extend their contribution to the literature in order to investigate the ability of the model regarding adversarial attack detection using deep learning techniques. Furthermore, it suggests suitable countermeasures for enhancing the ability of the model against the adversarial examples introduced in the training or testing phases.

Table 1 summarizes the existing studies and provides an analysis in terms of the mechanism for anomaly detection, dealing with adversarial attacks, and proposing a defense method.

Table 1. Summary of the existing studies.

Study	Providing Anomaly Detection	Dealing with Adversarial Attacks	Proposing Defense Method
[4]	✓	✕	✕
[5]	✓	✕	✕
[8]	✓	✕	✕
[9]	✓	✕	✕
[10]	✓	✕	✕
[10,11]	✓	✕	✕
[12]	✓	✕	✕
[13]	✓	✕	✕
[14]	✓	✕	✕
[15]	✓	✓	✓
[16]	✓	✓	✕
[17]	✓	✓	✕
[18]	✓	✓	✕
[19]	✓	✓	✓
[20]	✓	✓	✕
[21]	✓	✕	✕

An investigation of the existing literature shows that in order to increase the efficacy of anomaly detection strategies in WBANs, several methodologies, including statistical and ML techniques, have been developed by researchers. However, these methods have some drawbacks. For instance, statistical approaches cannot handle the dynamic nature of WBANs, making it challenging to identify a suitable evaluation threshold value. In addition, due to their processing complexity, non-parametric statistical methods are unsuitable for real-time applications. Numerous contemporary methodologies have employed ML algorithms, including decision trees, linear regression, ANN, and the k-nearest neighbor and random forest algorithms. In a sensitive domain that needs great accuracy and good performance, such as in healthcare, better solutions may be found than the deployment of machine learning. These algorithms have drawbacks when dealing with complicated and large datasets, which slow down processing, and are not effective in anticipating new anomalous patterns.

Deep learning is recommended by the research community due to its better performance; it reduces the necessity of handcrafted feature engineering and increases performance compared with traditional machine learning. Evaluating the model's performance is considered the most significant step in deep learning development. However, adversarial attacks have recently been compromising its effectiveness and have led to misclassification in deep learning models, affecting the model's performance. After reviewing the published literature, it is clear that most studies have not fully investigated the scenario of adversarial attacks, such as in [4,5,7]. Although the machine and deep learning techniques used will encounter several types of possible security concerns in the various phases, one of the most pressing concerns is the adversarial attack, wherein the system's adversaries are highly motivated to change the outcomes of models or to source personal data for their own gain. For example, attackers can cause misclassification by manipulating the data samples. In addition, few studies have investigated adversarial attacks, such as in [17,18,20], but the authors have not proposed any defensive method by which to protect the models. Defense methods in intelligent networks aim to distinguish malicious activity from the common patterns in intelligent networks. To mitigate the different variants of adversarial attack, these defense methods must assure the security of the data that have been gathered by the networked devices to combat various adversarial attacks. Some of the existing studies propose defensive methods, such as those reported in [15,19]. However, the authors of [15] focused on one type of attack that occurred in the training data and ignored the attacks that occurred in the testing data. In [19], the study was limited to attacks being countered by traditional ML techniques; the model was designed to deal with batch learning, a system that is difficult to apply to big data scenarios, which are characteristic of today's healthcare monitoring applications. Therefore, the proposed research in this paper is motivated by the need to develop a deep learning model for processing big data scenarios and to propose a possible solution for adversarial attack issues in WBANs.

3. The Proposed Model

This paper presents a technique for deep learning-based anomaly detection for WBANs that affords them the ability to detect adversarial attacks. The suggested model involves four phases: the data collection and pre-processing phase, the training phase, the detection phase, and the evaluation phase. In the evaluation phase, we test the ability of the proposed model to detect the adversarial examples that have been introduced in either the training or testing phases, and we implement a defense method that can mitigate the adversarial attacks. Figure 1 depicts the phases of the suggested model and the workflow of the various processes.

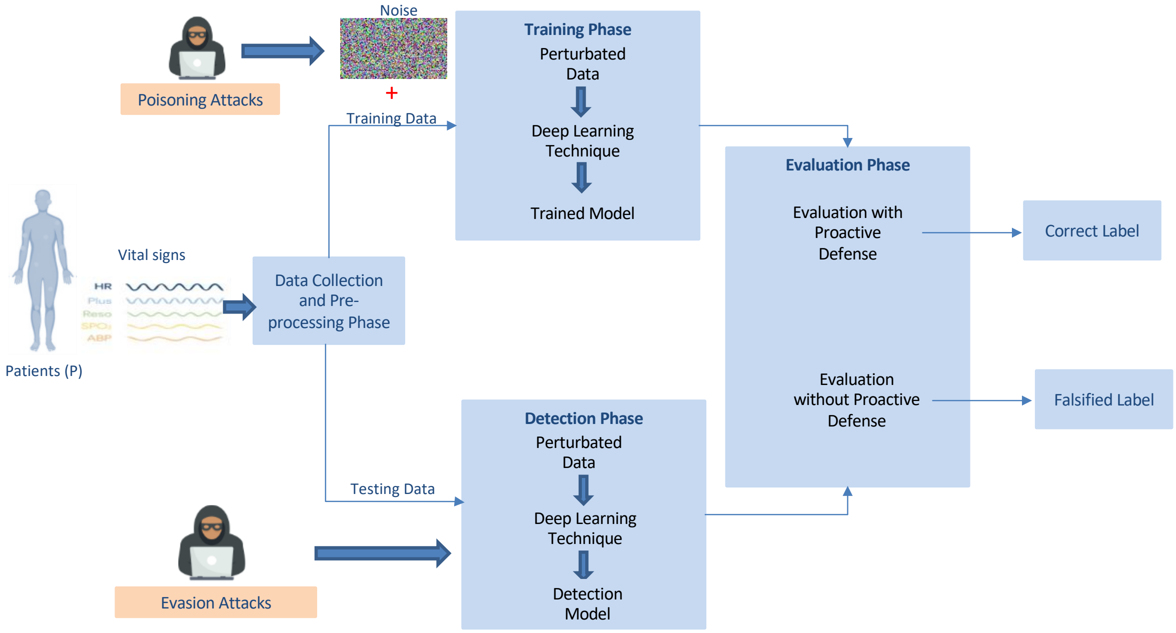


Figure 1. The Proposed Model.

3.1. Data Collection and Pre-Processing

In this phase, the data are collected from various WBAN sensors. The WBAN in question contains many wearable sensor nodes, which are used to collect physiological data including systolic arterial blood pressure (ABPsys), pulse rate, respiration rate (RESP), heart rate (HR), mean arterial blood pressure (ABP-mean), oxygen saturation (SPO_2), temperature, and diastolic arterial blood pressure (ABPdias), which are all measurements that should be taken into consideration when assessing a patient's condition. The Multiple Intelligent Monitoring in Intensive Care (MIMIC-I and II) dataset includes precise physiological data records taken from over 90 ICU patients, who are known as subjects. The dataset has seven features that represent the patient's clinical situation, which include mean arterial blood pressure (ABP-mean), oxygen saturation (SPO_2), heart rate (HR), temperature, diastolic arterial blood pressure (ABPdias), respiration rate (RESP), systolic arterial blood pressure (ABPsys), and pulse, with the appropriate timesteps and date; this dataset was used for the purposes of this study [22]. Table 2, below, contains a sample of the sensor readings in the dataset.

Table 2. The dataset's sample sensor readings.

Time and Date	HR	ABPSys	ABPDias	ABPMean	PULSE	RESP	SpO$_2$
14:07:00 10/11/15	77.6	157.4	66.1	100.5	77.9	23	97.4
14:08:00 10/11/15	77.3	149.2	62.6	95	77.6	22.2	97
14:09:00 10/11/15	76.1	150.5	62.4	95.1	76.8	22.3	97
14:10:00 10/11/15	73	158.4	65.4	99.8	74.3	22.2	97.4
14:11:00 10/11/15	75.6	152.4	63.3	96.7	76.4	22.4	97.5
14:12:00 10/11/15	75	154.3	63.4	97.1	75.4	22.2	97.5
14:13:00 10/11/15	75.2	150.3	62.1	94.7	76.7	22.1	97.6

The dataset is first preprocessed and prepared for use with the deep learning models via the normalization approach, which reorders the dataset's values appropriately. The goal of normalization is to use a consistent scale to reorder the values of the dataset's numeric columns without distorting the ranges of the values or erasing data. Additionally, certain algorithms require normalization in order to correctly simulate the data [23]. For the purposes of this paper, each attribute in the dataset is normalized, as in Equation (1), in the range [0, 1]:

$$x(i) = \frac{x(i) - \overline{x}}{S(x)} \tag{1}$$

where $x(i)$ is the dataset, $\overline{x}$ is one column in the dataset, and $S(x)$ is the number of data samples.

3.2. Training and Detection Using ConvLSTM

The training and detection phases are necessary to create a model in the training phase that can be used in the detection phase to identify anomalies in the WBAN, utilizing a deep learning approach. This technique has been widely implemented in healthcare applications because of its capacity to automatically detect complicated features without requiring domain expertise. The model developed in this paper is based on the hybrid ConvLSTM technique, which is a hybridization of the long short-term memory model and convolutional deep neural networks.

(a) Convolutional LSTM (ConvLSTM)

The hybrid deep learning system developed in this study combines the principles of convolutional and LSTM models. ConvLSTM is a type of RNN that is capable of capturing spatial-temporal data [24]. The convolutional component acquires the spatial area data, while the LSTM component leverages the temporal area data. However, the data, which are in the form of time series from the different sensors, have both spatial and temporal links. Thus, ConvLSTM can be employed as an important framework for anomaly detection problems involving time series data [25]. The ConvLSTM captures and applies both temporal and spatial correlations to predict the future state of a network cell, based on the inputs and previous states of its nearest neighbors. This is accomplished by including a convolution operation into the matrix multiplication step that is used in traditional fully integrated LSTM state-to-state and input-to-state transitions [24]. The ConvLSTM consists of multiple gates; the data flow may be stated using Equations (2)–(7).

The weighted sum of each gate's inputs is given a sigmoid function, which is then used as the activation function [26]:

$$f_t = \sigma\Big(W_{xf} \bigotimes X_t + W_{hf} \bigotimes H_{t-1} + W_{cf} \bigodot C_{t-1} + B_f\Big) \tag{2}$$

where f_t is the forget gate's output, which regulates the data that are lost in the previous cell state. C_{t-1}, W_{xi} is the convolution kernel used for the input tensor X_t in the input gate, X_t is the input tensor at time t, W_{hf} is the convolution kernel used for the input tensor H_{t-1} in the forget gate, H_{t-1} is the output tensor from the cell at time $t-1$, W_{cf} is the weight that is used for the old cell state C_{t-1} in the forget gate, C_{t-1} is the cell state at time $t-1$, and B_f is the bias in the forget gate.

$$\imath_t = \sigma\Big(W_{xi} \bigotimes X_t + W_{hi} \bigotimes H_{t-1} + W_{ci} \bigodot C_{t-1} + B_i\Big) \tag{3}$$

Here, $\imath_t$ is the output of the input gate, W_{xi} is the convolution kernel used for the input tensor X_t in the input gate, W_{hi} is the convolution kernel used for the input tensor H_{t-1} in the input gate, W_{ci} is the weight that is used for the old cell state C_{t-1} in the input gate, and B_i is the bias in the input gate.

$$c'_t = \tanh\Big(W_{xc} \bigotimes X_t + W_{hc} \bigotimes H_{t-1} + B_c\Big) \tag{4}$$

Here, c'_t is the data that are stored in the new cell state, C_t, W_{xc} is the convolution kernel used for the input tensor X_t to create the data $c\prime_t$ that will be stored in the new cell state c_t, W_{hc} is the convolution kernel used for the input tensor H_{t-1} to create the data $c\prime_t$ that will be kept in the new cell state c_t, and B_c is the bias for forming the data $c\prime_t$ that will be kept in the new cell state c_t.

$$c_t = f_t \bigodot C_{t-1} + c\prime_t \tag{5}$$

Here, c_t is the cell state at time t, and $c\prime_t$ is the data that are stored in the new cell state C_t.

$$o_t = \sigma\Big(W_{xo} \bigotimes X_t + W_{ho} \bigotimes H_{t-1} + W_{co} \bigodot C_t + B_o\Big) \tag{6}$$

Here, o_t is the output of the output gate; this controls the data that are output as h_t from the cell. W_{xo} is the convolution kernel used for the input tensor X_t in the output gate, W_{ho} is the convolution kernel used for the input tensor H_{t-1} in the output gate, W_{co} is the weight that is used for the new cell state c_t in the output gate, and B_o is the bias in the output gate.

$$h_t = o_t \bigodot \tanh(c_t) \tag{7}$$

Here, h_t is the output tensor from the cell at time t.

3.3. Adversarial Attack Modeling

Two types of adversarial attacks are introduced in the training and testing phases to examine the performance of the suggested model against such attacks. Poisoning adversarial attacks are introduced in the training phase, whereas evasion-based adversarial attacks are introduced in the testing phase. This adversarial attack can be defined as an input engineered to cause misclassification in the DL algorithms. Recently, adversarial DL has gained great popularity in healthcare applications, due to the limitations of the current DL models. For example, an adversary might insert new adversarial data into a healthcare DL model to falsely classify a hypothyroid patient [27]. Furthermore, in the context of medical image processing, the researchers reported various adversarial attacks against the DL model that were intended to modify results by inserting noise and causing the misclassification of a benign mole as malignant with high confidence [28,29]. In the previous section, the DL model was applied to detect normal and abnormal behaviors regarding the patient's vital signs. The following sections describe the types of adversarial attacks that have been adopted in this work.

(a) Poisoning Attack: This type of attack occurs during the training phase of the proposed model, as shown in Figure 1. In this attack, an adversary manipulates the training data to compromise the entire learning process. Modification, logic corruption, and data injection techniques can be used by an adversary to manipulate the training data. These capabilities allow an adversary to influence the proposed model's overall learning process and cause it to misclassify test results, which might result in the inappropriate treatment of a patient.

(b) Evasion Attack: In evasion attacks, the adversary attempts to deceive the model by mounting adversarial attacks during the testing phase, as shown in Figure 1. Such an adversary does not impact the training data, but they may access the proposed model to gain sufficient information. Consequently, they attack the proposed model and manipulate it so that it will misclassify the patient's condition. An adversary may use the fast gradient sign method (FGSM) attack or the basic iterative method attack (BIM) to perform evasion attacks.

(c) Fast gradient sign method of attack (FGSM): This method uses the gradient of the underlying model. The original input is manipulated by subtracting or adding a small error in the direction of the gradient, with the intent of altering the behavior of the learning model.

(d) Basic iterative method of attack (BIM): The BIM is an extension of the FGSM attack. The BIM attacks are repeated multiple times, using a small size, and are clipped. After each iteration, the result is clipped to guarantee the level of perturbation [30]. Both FGSM and BIM attacks are used to evaluate the robustness of the proposed model in this paper.

3.4. Proactive Defense

A proactive defense method is applied to counter adversarial attacks and to improve the performance of the proposed model. The applied defense comprises the retraining of the deep learning model (in the detection phase), which is based on robust optimization and model parameters to mitigate the effects of adversarial attacks that can cause higher confidence and distortion.

3.5. Model Evaluation

The suggested model's performance was evaluated using four metrics, namely, recall, accuracy, precision, and F1-score.

The class of each data object is established using specific classification methods, providing each sample with a predicted label (positive or negative) [31]. As a result, after the detection phase, each sample falls into one of four categories:

- True positives are actual positives, including the number of cases that are correctly identified as anomalous by the model (TP).
- False negatives are positives, including the number of cases that are misclassified as non-anomalous by the model (FN).
- True negatives are actual negatives, including the number of cases that are correctly classified as non-anomalous by the model (TN).
- False positives are actual negatives, including the number of cases that are misclassified as anomalous by the model (FP).

Four parameters, comprising recognition precision, recall, F1-score, and accuracy, are utilized to evaluate the effectiveness of the suggested model, utilizing the four classes listed above.

The accuracy value is a statistical measure of how efficiently a model can predict an outcome [32]. Equation (8) illustrates how the accuracy measure is calculated.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{8}$$

Recall and precision are frequently utilized to evaluate a result's correctness [33]; these are correctly expressed in Equations (9) and (10).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

The F1-score is the weighted sum of recall and precision and is utilized when the data are uneven [34]; Equation (11) illustrates how the F1-score is calculated.

$$\text{F1} - \text{score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{11}$$

4. Experiments and Results

In this section, the implementation of the suggested model and the investigation of its performance in relation to adversarial attacks are discussed. The section starts by providing the model parameters and the structure of the ConvLSTM technique.

4.1. Model Setup

This paper selects the MIMIC dataset [22], a substantial physiological dataset that has been referred to in Section 3.1. The suggested model was put into practice via Python, using the sklearn library with help from additional scientific computing libraries, such as Matplotlib, NumPy, and scikit-learn, to carry out a variety of tasks that include pre-processing and model selection. The ConvLSTM approach uses the Adam optimizer as its optimization algorithm. The ConvLSTM has a four-layer network with two dropout layers ((filters = 64 in two layers), (kernel_size = 1), (padding = "same" in two layers), and (activation function = "Relu")) with a kernel size of 1. The model was trained with a batch size of 30 and using various epoch counts. Using a ratio of 70:30 for the training and testing partitions, the dataset was divided into training and testing datasets. Due to the fact that the dataset is huge, this split is more suitable for this particular dataset. Table 3 summarizes the basic settings of the model.

Table 3. Model setup.

	Value
Language	Python
Libraries	pandas, NumPy, scikit-learn, Matplotlib, and Keras
Training set	70%
Test set	30%
Layers	4
Kernel_size	1
Activation function	Relu
Filters	64
Dropout	2
Optimizer	Adam
Number of epochs	30

4.2. Results and Analysis

In this subsection, we investigate the effect of the two types of adversarial attacks on the performance of the suggested anomaly detection model. Furthermore, we investigate how proactive defense can play a role in combating such adversaries.

4.2.1. The Results of a Poisoning Attack

In this type of attack, the attack poisons the training data in order to access the proposed model, so as to manipulate the output status of the patient and alter abnormal behaviors to normal behaviors, which will affect the patient's status. The attacker uses data modification, which poisons the training data directly by modifying them before they are used for training, in order to manipulate the model's decision (this is referred to as Data Poisoning Attack 1), as shown in Table 4. Another type of poisoning attack entails logic corruption, which affects the overall learning process of the proposed model (this is referred to as Data Poisoning Attack 2), as shown in Table 5. The results in Tables 4 and 5 show a decrease in the detection accuracy, as well as in the other performance metrics. Two data subjects, namely, subject 1 and subject 2, were examined. Table 4 shows that the accuracy of the model regarding subject 1, without the attacks, was 97%. This decreased to 74% after employing Data Poisoning Attack 1. For subject 2, the model without attacks achieved an accuracy of 99.9%; after applying Data Poisoning Attack 1, the accuracy dropped to 75.59%. Similarly, Table 5 reports the results of Data Poisoning Attack 2, which also shows a drop in all performance metrics. It reveals that the accuracy drops to 59% for subject 1 and to 75% for subject 2. As a result, the poisoning attacks were successful in compromising the proposed model's performance and causing data misclassification, which would impact the status of both patients and their treatments.

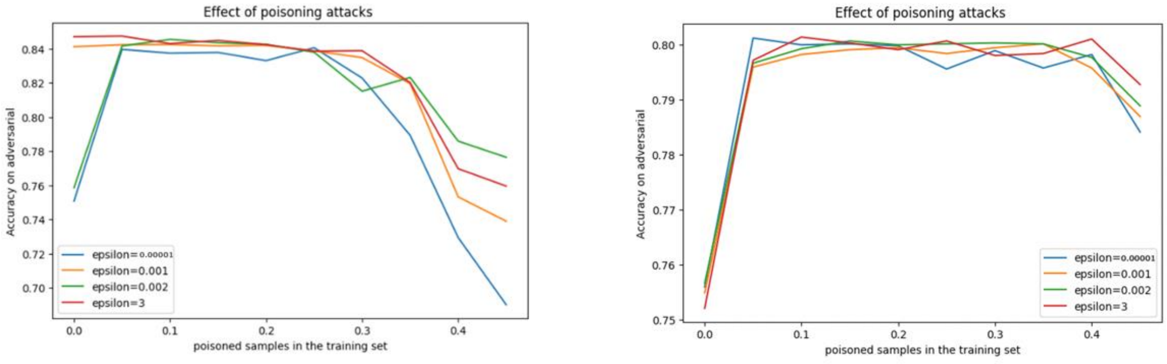
Table 4. Data Poisoning Attack 1 results.

Subject No.	With Attack or Without	Accuracy	Loss	Recall	Precision	F1-Score
Subject 1	Before Attack	96%	0.1%	99%	95%	97%
	Data Poisoning Attack 1	74%	0.4%	63%	40%	49%
Subject 2	Before Attack	99.91%	0.01%	99.93%	99.87%	99.90%
	Data Poisoning Attack 1	75.59%	0.24%	57%	76%	41%

Table 5. Data Poisoning Attack 2 results.

Subject No.	With Attack or Without	Accuracy	Loss	Recall	Precision	F1-Score
Subject 1	Before Attack	97%	0.1%	99%	95%	97%
	Data Poisoning Attack 2	59%	0.5%	39%	54%	68%
Subject 2	Before Attack	99.9%	0.01%	99.9%	99.87%	99.9%
	Data Poisoning Attack 2	75%	0.18%	55%	54%	53%

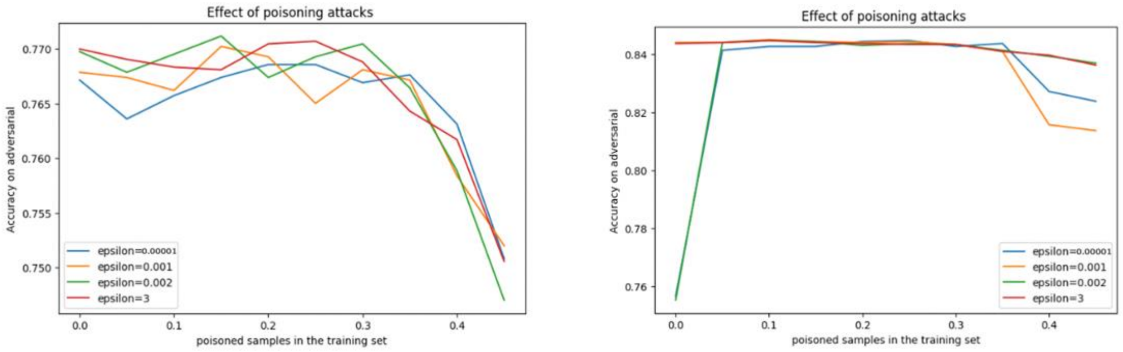
In Figure 2, different epsilon values, which represent the amount of adversarial content in the data, are used to test the efficacy of the proposed model against the adversarial Data Poisoning Attack 1. In the dataset for subject 1, various values of epsilon (0.00001, 0.00100, 0.00200, and 3.00000) were examined and resulted in different accuracy values (67%, 75%, 78%, and 75%). Similarly, for subject 2, the same epsilon values were examined and resulted in identical accuracy values (79%, 79%, 79%, and 79%).



(**a**) (**b**)

Figure 2. The variations in the adversarial Data Poisoning Attack 1 results (epsilon): (**a**) subject 1, (**b**) subject 2.

The same analogy applies to adversarial Data Poisoning Attack 2. As shown in Figure 3, the epsilon values were used for subject 1 and achieved different accuracy values of (75%, 75%, 75%, and 75%) and (82%, 81%, 84%, and 84%) for subject 2.



(**a**) (**b**)

Figure 3. The variations in the data from adversarial Data Poisoning Attack 2 (epsilon): (**a**) subject 1, (**b**) subject 2.

The results reported in Figures 2 and 3 clearly show that as the amount of adversarial content increases (the epsilon values increase), the accuracy decreases; it drops sharply after epsilon = 0.3. This outcome can be generalized for both data subjects and for both poisoning attack types.

4.2.2. The Results of Evasion Attacks

For evasion attacks, the attacks adjusted the data to introduce malicious samples during the testing phase. This was achieved by performing gradient-based operations to fool the DL model, stopping it from detecting any strange behavior. As a result, a drop in the accuracy of the model and in other performance metrics was recorded. The BIM and FGSM attacks affected the performance of the model for all datasets and fooled the DL model, as shown in Tables 6 and 7. In the case of subject 1 and subject 2, we created two scenarios; these are named "Before Attack" and "After BIM Attack", as shown in Table 6. For both scenarios, we used the same parameters to report the results. In the case of subject 1, Table 6 shows that before the attacks, the achieved accuracy was 97.59%, with a loss of 0.1%. However, a severe accuracy drop to 79% was reported after the BIM attack. A similar scenario can be seen for subject 2, showing a similar drop in all performance metrics.

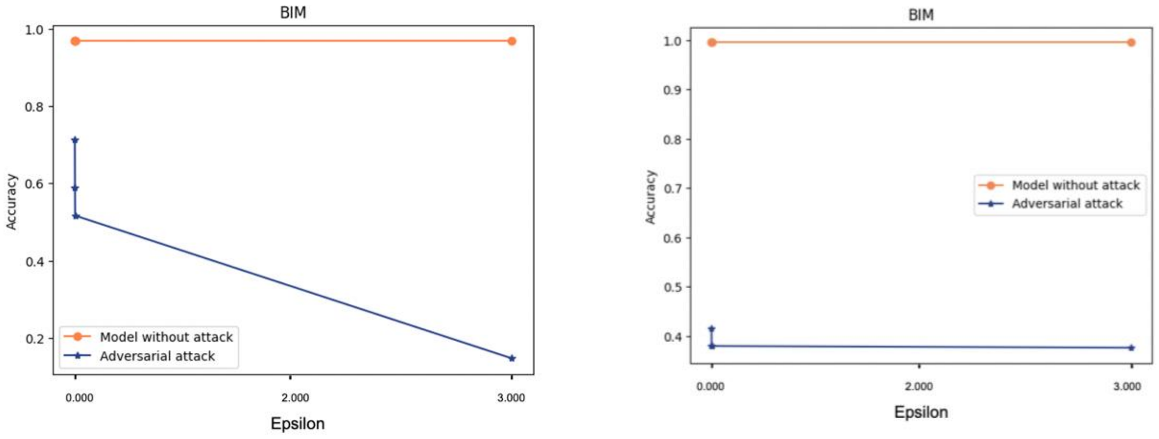
Table 6. The BIM attack results.

Subject No.	With Attack or Without	Accuracy	Loss	Recall	Precision	F1-Score
Subject 1	Before Attack	97.59%	0.1%	99.40%	95.56%	97.59%
	After BIM Attack	79%	0.4%	50%	31%	73%
Subject 2	Before Attack	99.99%	0.01%	99.93%	99.87%	99.90%
	After BIM Attack	55%	0.2%	51%	78%	37%

Table 7. FGSM attack results.

Subject No.	With Attack or Without	Accuracy	Loss	Recall	Precision	F1-Score
Subject 1	Before Attack	97%	0.1%	99%	95%	97%
	After FGSM Attack	63%	0.5%	50%	50%	33%
Subject 2	Before Attack	99%	0.01%	99.93%	99.87%	99.90%
	After FGSM Attack	71%	0.3%	53%	49%	53%

As seen in Figure 4a, we applied different epsilon values, which represent the amounts of BIM adversarial attack content for the subject 1 dataset (0.00001, 0.00100, 0.00200, and 3.00000). The figure shows the achieved accuracy values of 71%, 59%, 52%, and 15%, respectively. In Figure 4b, we applied the same epsilon values, which resulted in different accuracy values of 42%, 38%, 38%, and 38%, respectively. In Figure 4, we applied a different approach by depicting the accuracy value before and after the attacks, along with the variations in the epsilon values. It is clear that a sharp drop was recorded, which increased along with the increase in the amount of adversarial content.



(**a**) (**b**)

Figure 4. The variations in adversarial BIM attack content (epsilon): (**a**) subject 1, (**b**) subject 2.

In a similar way, Table 7 shows the results of the proposed model before and after the FGSM attacks, which can be considered a type of evasion attack. As is reported, for both subjects, the model exhibited a sharp drop in all metrics after the attacks.

Figure 5a shows the similarly applied different epsilon values, which represent the amounts of FGSM adversarial attack content for the subject 1 dataset (0.00001, 0.00100, 0.00200, and 3.00000), and shows the achieved accuracy values of 63%, 63%, 63%, and 61%, respectively. Figure 5b shows the similarly applied same epsilon values, which resulted in different accuracy values of 71%, 71%, 71%, and 69%, respectively. It is clear that a sharp drop has been recorded, which increased along with the increase in the amount of adversarial content for subject 1 but exhibited a slight increase in accuracy for subject 2. The reason behind such unusual behavior in the model regarding subject 2 might be related to the behavior of the model without an attack regarding that subject, which also exhibited a decrease in accuracy.

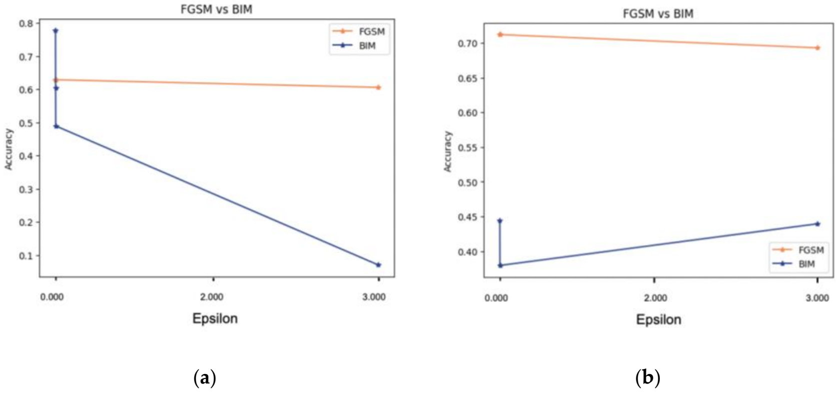


Figure 5. The variations in adversarial FGSM attack content (epsilon): (**a**) subject 1, (**b**) subject 2.

In summary, the investigation of two types of attacks showed that all of the attacks decreased the performance of the model. The BIM attack achieved the highest decrease compared to the rest of the attacks, as shown in Figures 4 and 5.

4.2.3. Proactive Defense Method

Retraining the model is a common training practice used to reduce losses and increase its accuracy. This is one of the most effective proactive strategies for defense against adversarial attacks on deep learning models. The resultant retrained network should be able to withstand the adversarial attacks used to generate adversarial samples during the training phase. These adversarial attacks target neural networks to cause misclassification. After the attacks, these assaults and the correct labels are used to teach the neural network. Such proactive strategies cause a minor decline in the predictability of a deep learning model. However, the resilience that they provide against adversarial attacks is thought to make up for this decline. A model of this type of operation is illustrated in Figure 6, below.

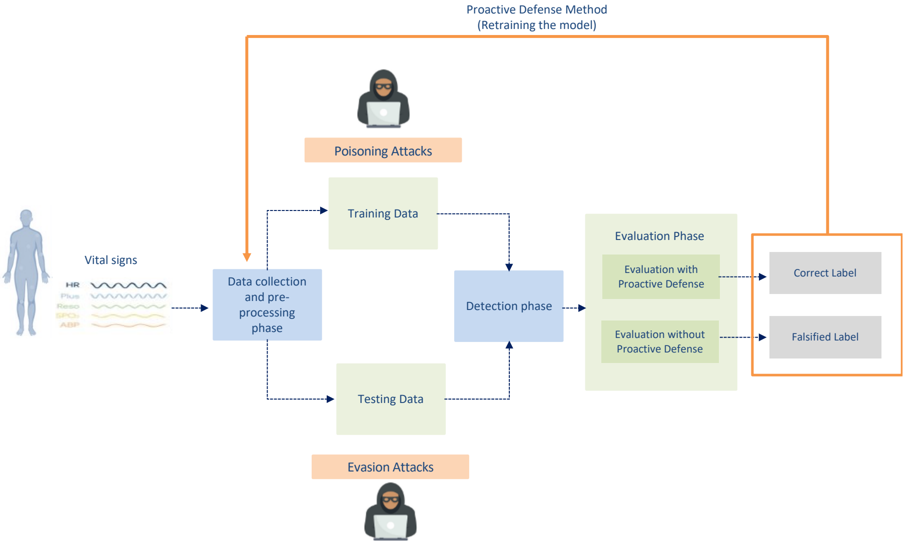


Figure 6. The proactive defense method.

Figure 7 and Table 8 show an increase in accuracy, as well as in other performance metrics, after retraining the model for both subjects.

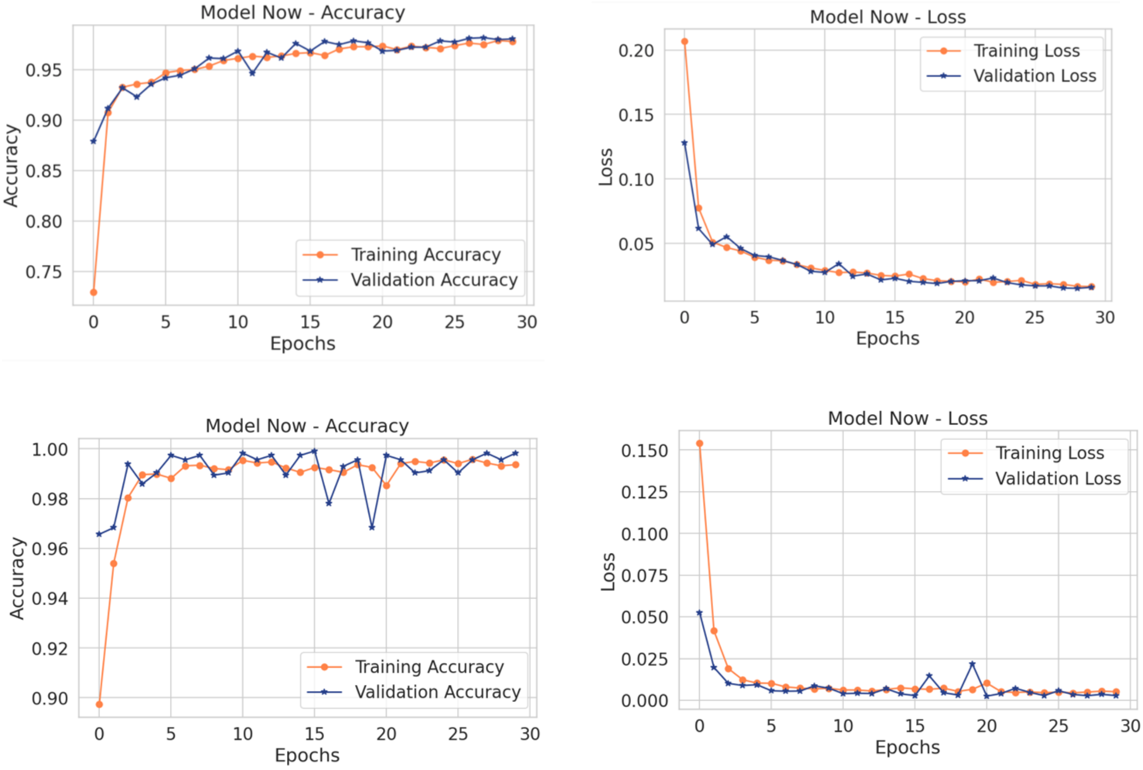


Figure 7. The model after retraining as a defense method.

Table 8. Retraining the model as a defense method.

Subject No.	Accuracy	Loss	Recall	Precision	F1-Score
Subject 1	98%	0.05%	98%	97%	97%
Subject 2	99.36%	0.005%	100%	100%	100%

Analyzing the scenarios where both attacks succeeded in fooling the DL model shows that deep learning models have weaknesses that may lead to misclassifying normal behavior as abnormal behavior. In the healthcare domain, our investigations revealed that adversarial attacks can reduce accuracy when utilizing the most recent deep learning model and, therefore, lead to severe and dangerous consequences. Thus, retraining the model is a proactive defensive method that helps to make sure that such attacks fail to fool the deep learning model and makes the model more robust.

5. Conclusions

With the continuous advancement of smart sensors, wireless communications, and IoT applications and services, wireless body area networks (WBANs) are becoming increasingly popular. However, several adversarial attacks and the attendant weaknesses make it challenging to develop secure IoHT applications from a security perspective. The widespread usage of anomaly detection with deep learning has been employed to construct appropriate models for detecting and mitigating adversarial attacks. Nonetheless, these attacks could deceive deep learning models into misclassifying the model results. This can lead to erroneous decisions, such as incorrect patient diagnoses and erroneous medicine administration. This paper addressed the effect of adversarial attacks on deep learning-based anomaly detection models and applied several safeguard techniques to defend networks against such adversarial attacks. The effectiveness of the suggested model was assessed in the face of adversarial attacks that were intended to give attackers control over the classification process used by the anomaly detection algorithm. The effectiveness of the suggested model was evaluated using the FGSM and BIM adversarial attack techniques. The suggested model showed its capacity to identify anomalies in the face of various adversarial attacks. More specifically, the evaluation results revealed that the model achieved an average F1 score of around 97% and an accuracy of 98% in the face of such attacks. However, as a future task, more scenarios regarding adversarial attacks need to be explored. In addition, more countermeasures and defense methods, such as reactive defense, should be investigated.

Author Contributions: Conceptualization, A.A. and M.A.R.; formal analysis, A.A. and M.A.R.; investigation, A.A. and M.A.R.; methodology, M.A.R.; software, A.A.; validation, A.A.; writing—original draft, A.A.; writing—review and editing, M.A.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research has not received funding.

Acknowledgments: The authors gratefully acknowledge Qassim University, represented by the Deanship of Scientific Research, on the financial support for this research under the number (COC-2022-1-1-J-25671) during the academic year 1444 AH/2022 AD.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhou, X.; Liang, W.; Li, W.; Yan, K.; Shimizu, S.; Wang, K.I.-K. Hierarchical Adversarial Attacks Against Graph-Neural-Network-Based IoT Network Intrusion Detection System. *IEEE Internet Things J.* **2021**, *9*, 9310–9319. [CrossRef]
2. Pachauri, G.; Sharma, S. Anomaly Detection in Medical Wireless Sensor Networks using Machine Learning Algorithms. *Procedia Comput. Sci.* **2015**, *70*, 325–333. [CrossRef]
3. West, J.; Bhattacharya, M. Intelligent financial fraud detection: A comprehensive review. *Comput. Secur.* **2016**, *57*, 47–66. [CrossRef]
4. Dehabadi, M.S.Z.; Jahed, M. Reliability modeling of anomaly detection algorithms for Wireless Body Area Networks. In Proceedings of the 2017 Iranian Conference on Electrical Engineering (ICEE), Tehran, Iran, 2–4 May 2017; pp. 70–75. [CrossRef]
5. Saneja, B.; Rani, R. An integrated framework for anomaly detection in big data of medical wireless sensors. *Mod. Phys. Lett. B* **2018**, *32*, 1850283. [CrossRef]
6. Ibitoye, O.; Shafiq, O.; Matrawy, A. Analyzing Adversarial Attacks against Deep Learning for Intrusion Detection in IoT Networks. In Proceedings of the 2019 IEEE Global Communications Conference (GLOBECOM), Waikoloa, HI, USA, 9–13 December 2019; pp. 1–6. [CrossRef]
7. Salem, O.; Serhrouchni, A.; Mehaoua, A.; Boutaba, R. Event Detection in Wireless Body Area Networks Using Kalman Filter and Power Divergence. *IEEE Trans. Netw. Serv. Manag.* **2018**, *15*, 1018–1034. [CrossRef]
8. Al Rasyid, M.U.H.; Setiawan, F.; Nadhori, I.U.; Sudarsonc, A.; Tamami, N. Anomalous Data Detection in WBAN Measurements. In Proceedings of the 2018 International Electronics Symposium on Knowledge Creation and Intelligent Computing (IES-KCIC), Bali, Indonesia, 29–30 October 2018; pp. 303–309. [CrossRef]
9. Mohamed, M.B.; Makhlouf, A.M.; Fakhfakh, A. Correlation for efficient anomaly detection in medical environment. In Proceedings of the 2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC), Limassol, Cyprus, 25–29 June 2018; pp. 548–553. [CrossRef]
10. Nair, S.G.S.; Balakrishnan, R. Mitigating false alarms using accumulator rule and dynamic sliding window in wireless body area. *CSI Trans. ICT* **2018**, *6*, 203–208. [CrossRef]
11. Smrithy, G.S.; Balakrishnan, R.; Sivakumar, N. Anomaly Detection Using Dynamic Sliding Window in Wireless Body Area Networks. In *Data Science and Big Data Analytics: ACM-WIR 2018*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 99–108. [CrossRef]
12. Arfaoui, A.; Kribeche, A.; Senouci, S.M.; Hamdi, M. Game-based adaptive anomaly detection in wireless body area networks. *Comput. Netw.* **2019**, *163*, 106870. [CrossRef]
13. Nagdeo, S.K.; Mahapatro, J. Wireless Body Area Network Sensor Faults and Anomalous Data Detection and Classification using Machine Learning. In Proceedings of the 2019 IEEE Bombay Section Signature Conference (IBSSC), Mumbai, India, 26–28 July 2019; pp. 1–6. [CrossRef]
14. Boudargham, N.; El Sibai, R.; Abdo, J.B.; Demerjian, J.; Guyeux, C.; Makhoul, A. Toward fast and accurate emergency cases detection in BSNs. *IET Wirel. Sens. Syst.* **2020**, *10*, 47–60. [CrossRef]
15. Qiu, H.; Dong, T.; Zhang, T.; Lu, J.; Memmi, G.; Qiu, M. Adversarial Attacks Against Network Intrusion Detection in IoT Systems. *IEEE Internet Things J.* **2020**, *8*, 10327–10335. [CrossRef]

16. Bovenzi, G.; Foggia, A.; Santella, S.; Testa, A.; Persico, V.; Pescape, A. Data Poisoning Attacks against Autoencoder-based Anomaly Detection Models: A Robustness Analysis. In Proceedings of the ICC 2022—IEEE International Conference on Communications, Seoul, Republic of Korea, 16–20 May 2022; pp. 5427–5432. [CrossRef]
17. Watson, M.; Al Moubayed, N. Attack-agnostic adversarial detection on medical data using explainable machine learning. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 5427–5432.
18. Raza, A.; Li, S.; Tran, K.P.; Koehl, L. Detection of Poisoning Attacks with Anomaly Detection in Federated Learning for Healthcare Applications: A Machine Learning Approach. *arXiv* **2022**, arXiv:2207.08486. [CrossRef]
19. Newaz, A.I.; Haque, N.I.; Sikder, A.K.; Rahman, M.A.; Uluagac, A.S. Adversarial Attacks to Machine Learning-Based Smart Healthcare Systems. In Proceedings of the GLOBECOM 2020–2020 IEEE Global Communications Conference, Taipei, Taiwan, 7–11 December 2020; pp. 1–6. [CrossRef]
20. AlZubi, A.A.; Al-Maitah, M.; Alarifi, A. Cyber-attack detection in healthcare using cyber-physical system and machine learning techniques. *Soft Comput.* **2021**, *25*, 12319–12332. [CrossRef]
21. Albattah, A.; Rassam, M.A. A Correlation-Based Anomaly Detection Model for Wireless Body Area Networks Using Convolutional Long Short-Term Memory Neural Network. *Sensors* **2022**, *22*, 1951. [CrossRef]
22. MIMIC2 Dataset. Available online: https://physionet.org/content/mimicdb/1.0.0/ (accessed on 22 October 2022).
23. Alghofaili, Y.; Albattah, A.; Rassam, M.A. A Financial Fraud Detection Model Based on LSTM Deep Learning Technique. *J. Appl. Secur. Res.* **2020**, *15*, 498–516. [CrossRef]
24. Shi, X.; Chen, Z.; Wang, H.; Yeung, D.Y.; Wong, W.K.; Woo, W.C. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. [CrossRef]
25. Shastri, S.; Singh, K.; Kumar, S.; Kour, P.; Mansotra, V. Time series forecasting of COVID-19 using deep learning models: India-USA comparative case study. *Chaos Solitons Fractals* **2020**, *140*, 110227. [CrossRef] [PubMed]
26. Xiao, C.; Chen, N.; Hu, C.; Wang, K.; Xu, Z.; Cai, Y.; Xu, L.; Chen, Z.; Gong, J. A spatiotemporal deep learning model for sea surface temperature field prediction using time-series satellite data. *Environ. Model. Softw.* **2019**, *120*, 104502. [CrossRef]
27. Mozaffari-Kermani, M.; Sur-Kolay, S.; Raghunathan, A.; Jha, N.K. Systematic Poisoning Attacks on and Defenses for Machine Learning in Healthcare. *IEEE J. Biomed. Health Inform.* **2014**, *19*, 1893–1905. [CrossRef]
28. Finlayson, S.G.; Bowers, J.D.; Ito, J.; Zittrain, J.L.; Beam, A.L.; Kohane, I.S. Adversarial attacks on medical machine learning. *Science* **2019**, *363*, 1287–1289. [CrossRef]
29. Newaz, A.I.; Sikder, A.K.; Rahman, M.A.; Uluagac, A.S. A Survey on Security and Privacy Issues in Modern Healthcare Systems: Attacks and defenses. *ACM Trans. Comput. Health* **2021**, *2*, 1–44. [CrossRef]
30. Fawaz, H.I.; Forestier, G.; Weber, J.; Idoumghar, L.; Muller, P.-A. Adversarial Attacks on Deep Neural Networks for Time Series Classification. In Proceedings of the 2019 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 14–19 July 2019; pp. 1–8. [CrossRef]
31. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genom.* **2020**, *21*, 6. [CrossRef]
32. Ahmed, M.; Mahmood, A.N.; Islam, R. A survey of anomaly detection techniques in financial domain. *Futur. Gener. Comput. Syst.* **2016**, *55*, 278–288. [CrossRef]
33. Lee, T.J.; Gottschlich, J.; Tatbul, N.; Metcalf, E.; Zdonik, S. Precision and recall for range-based anomaly detection. *arXiv* **2018**, arXiv:1801.03175. [CrossRef]
34. Singh, S.P.; Sharma, M.K.; Lay-Ekuakille, A.; Gangwar, D.; Gupta, S. Deep ConvLSTM with Self-Attention for Human Activity Decoding Using Wearable Sensors. *IEEE Sens. J.* **2020**, *21*, 8575–8582. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.