iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.52084

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

A Multi Model Approach: A Data Engineering Driven Pipeline Model for Detecting Anomaly in Sensor Data using Stacked LSTM

Subha[1], Anurag patel[2], Saranraj[3]
Department of Computer Science and Engineering, SRM Institute of Science and Technology, Ramapuram, Chennai, India

Abstract: ***This study uses a self-supervised learning technique based on auto encoders to find anomalous nodes. Only temporal variables have been taken into account and researched so far for use in identifying anomalies in wireless sensor networks (WSNs). This method fully utilises the geographic and temporal information of the WSN for anomaly identification by incorporating the extraction of geographic location features, intermodal WSN correlation features, and temporal WSN data flow characteristics into the design of the autoencoder. First, by focusing on a single mode from a local spatial perspective, a fully connected network is used to temporal nodes. Second, the spatial and temporal characteristics of the data flows of the nodes and their neighbours are retrieved by concentrating on a specific mode and seeing the WSN topology from a global spatial perspective for anomaly identification. The adaptive fusion method's weighted summation step is then used to extract the relevant features from the various models. An LSTM is used in this study to solve the problem of long-term dependence in the temporal dimension. The decoder's reconstructed output and the hidden layer representation of the are used to calculate the anomaly probability of the current system utilising a fully linked network.***

Keywords: Anomely Detection (AD), stacked (LSTM)

I. INTRODUCTION

In unsupervised environments where WSNs are regularly used, node assaults are common. These attacks interfere with wireless media, intercept transmission messages, and degrade the deployment environment, among other things, preventing WSNs from operating normally. Anomalies happen when the WS Ns' energy memory bandwidth and communication limitations are violated. Contextual anomalies, point anomalies, and collective anomalies, to mention a few, are data flows that differ from the typical data distribution when an event occurs in a WSN. Particularly abnormal data points are those that differ in some way from the rest of the data. Data points that are contextual anomalies are out of the ordinary when contrasted to the regular data they are related with. Anomalies can be grouped together.. A single data point could seem to be normal, but if there are many of them at once, a larger anomaly is produced. In conclusion, anomalies might arise in a WSN at any time, so it's critical to spot them as soon as is practical and to halt degradation. The safe and dependable operation of WSN systems is implied to require anomaly detection. Sensor-based remote health monitoring is increasingly being used in a variety of urban, industrial, and healthcare situations. Knowledge derived from sensor-based data enables the analysis of temporal trends and the cheapest, most invasive diagnosis of serious diseases. Seniors' health and condition can be tracked using sensor data on things like movement, physiology, behaviour, and sleep. Seniors are able to maintain their independence for longer, and quick intervention is made possible. However, real-world sensor-based health monitoring presents unique challenges. Its distinguishing features include multivariate data, erroneous labelling brought on by resource-intensive annotation, data drift noise, and a lack of periodicity. lightweight self-supervised anomaly detection method that is robust against tagging and noisy data that are typical of sensor-based remote health monitoring. Our adaptive anomaly detection system for the healthcare sector uses specialist baseline data. Applying the Matrix Principle(MP) -a more adaptive, contemporary variation of the distance-based anomaly detection method, the Contextual Matrix Prole (CMP), which is modern, accurate, and remarkably quick. The CMP concept's intuitive representation of patterns and anomalies is its cornerstone. The CMP facilitates the distinction between normal and abnormal data by organising noisy multivariate sensor observations into time intervals or contexts.Cross-sensor correlations and high dimensions are currently unaddressed by existing work based on the CMP. These drawbacks are overcome by our method by utilising graph-based machine learning. We explicitly build time context graphs using CMP distance techniques, and we apply self-supervised graph models to assess each graph in relation to the prior graphs.

Then, each graph embedding is subjected to the sliding window method to look for spatiotemporal anomalies. The usage of MP, which internally uses the Fast Fourier Transform for distance computation, and one-hop graphs, which benefit from graph representation learning but with few parameters and low computational complexity, both increase the speed of our method. Our research extends past work that demonstrates the potency of the CMP-based approach for unsupervised anomaly discovery. Additionally, downstream algorithms are able to comprehend and distinguish between particular anomaly conditions using the generated embeddings by layering graph models on top of the CMP.Cyber Physical Systems (CPS) are like a certain kind of car. Recent hardware innovations have enabled access to computational power at previously unheard-of levels through the usage of modern processors and GPUs. Using the extensive automotive datasets that machine learning and artificial intelligence algorithms have access to, intelligent solutions based on this technology can be developed. These techniques can be used to identify the causes of both anomalous sensor activity and other sensor channel issues. Monitoring issues, predicting their occurrence, and providing guidance for decision-making can all help with the creation of maintenance programmes.In both the academic and business realms, predictive maintenance (PdM) is a notion that is gaining popularity.But PdM installation is costly and challenging. To breach the PdM domain, robust anomaly detection is necessary. AI-driven anomaly detection systems are used to identify events or observations that drastically deviate from the majority of the data and do not meet a predetermined description of typical behaviour.The Internet of Things (IoT) is a rapidly expanding network that connects devices through complex linkages to enable data collecting and exchange. As the number of IoT users and applications increases across multiple industries, new issues with security and privacy of devices in the IoT network are presented. One of the research areas in current IoT data analytics is finding aberrant data or outliers in data streams. Anomalies, often known as outliers, or unusual patterns or behaviours in data may indicate a problem or a rare occurrence. .Anomalies may result from errors or strange discoveries. If weaknesses are not uncovered and fixed, hostile attacks might bring down the entire Internet of Things network. When the IoT network is in use, rare observations are uncommon occurrences that may happen and may need to be monitored or reported. For quickly recognising and responding to issues, anomaly detection in the IoT is essential. Anomalies in sensor data can reveal faulty equipment, whereas anomalies in network traffic could reveal a cyberattack. Anomaly detection can also find fraud or other strange behaviour in financial transactions, as well as discover anomalous trends in business activity. The ability of machine learning algorithms to analyse huge data sets has made them popular for anomaly detection.

II. RELATED WORK

A. Reconstruction Error-based Methodologies

Generic normalcy feature learning frequently uses the variance between an input and the reconstructed output (for instance, mean squared error) as a gauge of abnormality in DAD. Due to their greater capacity for latent representation learning, the autoencoder (AE) and variational autoencoder (VAE) have both been extensively used. The aberrant samples cannot be successfully reproduced from lower-dimensional latent properties once a neural network has been trained using normal samples to limit its reconstruction error. By correctly selecting a threshold for the normal instances, anomalies can be detected from such situations since they contain more reconstruction faults. a new writing has just published.Recovery along Projection Paths (RAPP), a revolutionary DAD methodology that makes use of latent space reconstruction defects The latent features for each layer of the encoder are first acquired along the first forward path. The distinction in the second forward path's latent characteristics is employed as an anomaly score by reinputting the output into the encoder network. Generative adversarial networks (GANs) serve as the base of another branch. Modelling the distribution of normal samples is possible with a neural network built using the GAN architecture. In GAN approaches, the discriminator loss and reconstruction error are combined to score anomalies.GANomaly is taken into account for anomalous scoring, as well as the incorrect reconstruction of the bottleneck features. After that, the latent spac defines the score.The GAN-based DAD's unified design exemplifies how the ensemble of anomaly scores from several GAN variations significantly boosts detection performance. Previous research has demonstrated that latent reconstruction errors, discriminator loss, and an ensemble of reconstruction defects all improve AD performance.This can be explained by a variety of anomaly sources. In a manner similar to this, ambiguity sources boost the effectiveness of DAD methods. B. Difficulties in Identifying Abnormal Changes Deep ensembles, Monte-Carlo (MC) dropout methodology, and Bayesian deep learning are the only techniques for handling uncertainty in deep learning contexts [52]. The MC dropout technique, for instance, is used in applications for uncertainty-based DAD. In order to account for dropout, MC dropout is used in both the inference and training phases. Because of the probabilistic connections between the neurons, a single neural network can produce many different outputs. After deciding to withdraw out of the inference phase, This tactic aims to profit primarily from epistemic ambiguity.

Using the MC dropout method, the following examples show how to assess epistemic uncertainty: Using the Uber dataset and a confidence interval, [31] conducted research on a deep learning-based time-series prediction. AD was performed by triggering 70430 VOLUME 10, 2022 S. QAE When the observed value diverged from the 95% predicted range, Ryu et al.'s AA for AD of Multivariate Sensor Data raised an alarm. The anomaly score was weighted by uncertainty rather than the variation in reconstruction mistakes. In the field of medical imaging, uncertainty was used to identify diabetic retinopathy using fundus pictures.employing pixel-level adjustments to the retinal optical coherence tomography images for segmentation. The uncertainty of aberrant pictures in the MVTec-AD dataset was quantified by comparing the area under the receiver operating characteristic (AUROC) scores between the residual-based and uncertainty-based detection results. In this study, aleatoric uncertainty is incorporated by QAE, and the distribution of the sources used for anomaly scoring is done using the uncertainty term. Our method differs from the earlier ones in how uncertainty is measured (aleatoric uncertainty with multiple quantile regression) and how anomaly scoring is done (Mahalanobis distance-based anomaly score)

III. FEASIBILITY STUDY

The viability of implementing the "Quantile Autoencoder for Anomaly Detection" project depends on a number of variables, including resource availability, data quality and quantity, and team technical knowledge.

These are some crucial ideas to bear in mind:

Resources: The QAE model for anomaly detection calls for a substantial amount of computational resources, including high-performance computer systems and specialised hardware like GPUs. To make sure the project can be completed, the cost and availability of these resources should be carefully addressed.

The effectiveness of the suggested QAE-AA is tested using several real-world datasets. In four of the six datasets, QAE-AA achieved the greatest AUROC score . These test results demonstrate that the suggested methodology can enhance AD performance. Utilising adversarial examples and time series AD settings, the proposed QAE-AA framework can also be used in some benchmark datasets. In this sense, the goal of our ongoing study is to get past the limitations and boost AD performance even further. For instance, epistemic uncertainty or latent space defects can be added to improve AD performance on image data can be used as an additional method to judge how well anomaly detection models work. When the model is used in a production environment, it must also be maintained and tracked.

Overall, despite possible challenges, implementing the QAE model for anomaly detection is doable with the right equipment, information, and careful planning.

IV. EXISTING SYSTEM

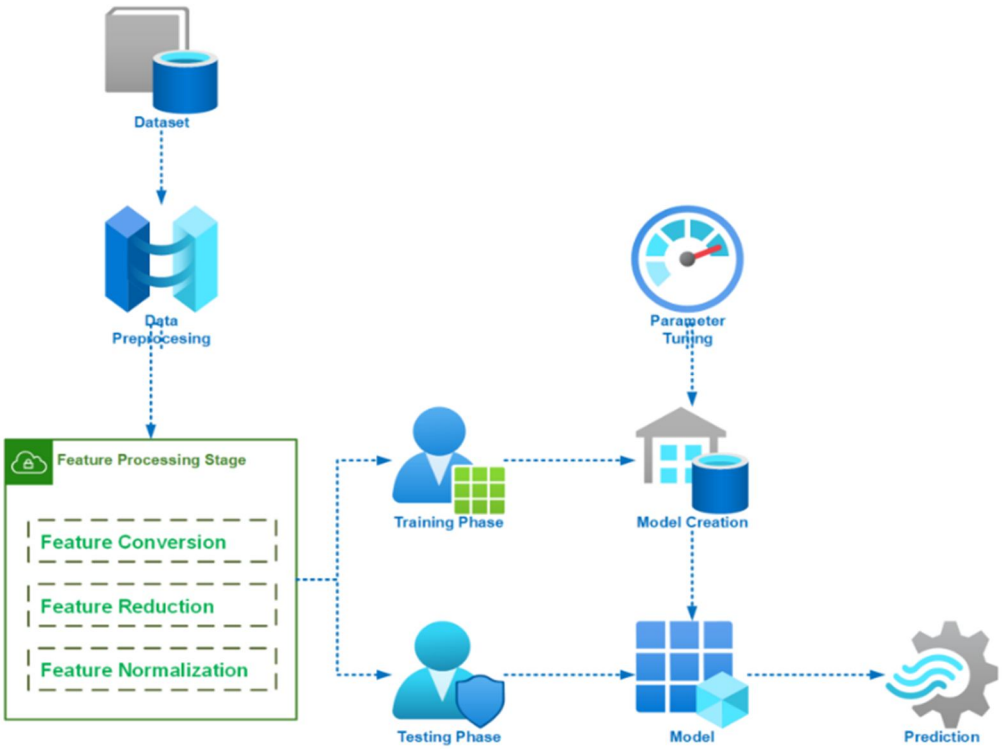
Anomaly detection (AD) is a crucial issue in many industrial settings where several sensors provide a large volume of data. Deep learning-based approaches for treating Alzheimer's disease (AD) have been significantly improved by big data and deep neural networks (DNN). Reconstruction error, which is the difference between the original and reconstructed values, is the metric of abnormality used by the majority of deep anomaly detection (DAD) systems. But by altering the source of the aberrant score, AD performance can be improved. We introduce the notion of anomaly source diversity and give mathematical justifications for it to support this. In response, we suggest the quantile autoencoder (QAE) with abnormality accumulation (AA), a revolutionary DAD technique, which accesses sources repeatedly. The uncertainty term—the range between the two quantiles—and the reconstruction error are both considered when calculating the anomaly score using QAE. Additionally, AA aggregates the mistakes discovered during the recursive reconstruction of the input before calculating the anomaly score using the Mahalanobis distance. By narrowing the distributions' breadth, this technique increases AD performance by increasing the distance between the score distributions of normal and abnormal data. The effectiveness of the proposed QAE-AA was investigated using tests on multi-variate sensor datasets from various domains; on average, QAE-AA outperforms previous AD approaches in terms of AUROC score. The QAE network and AA approach are suggested in this paper, which investigates the idea of anomaly score diversification. Experiments on actual datasets have confirmed the usefulness of the proposed framework. The benefit of varying anomaly sources Anomaly detection performance can be increased by reducing the reasons for anomaly score computation errors. By demonstrating how the distributions of mean square error anomaly scores on normal and abnormal diverge more as the number of error sources rises, we theoretically support this. This is accomplished by adding aleatoric uncertainty as an extra source of error for anomaly scoring, and by presenting a QAE that generates both the quantiles and the median. It is likely that outputs from anomalous samples will have higher channel-wise uncertainty than outputs from normal samples, similar to the reconstruction errors.

We also provide the AA method, which first uses Mahalanobis distance to determine anomaly score before accumulating errors over numerous reconstructions. As the recursion increases the dimension of the errors, it is easier to distinguish between the anomaly score distributions of normal and abnormal data. Several real-world datasets are used to assess the effectiveness of the proposed QAE-AA. QAE-AA achieved the highest AUROC scores and, on average, % to % higher AUROC scores in four of the six datasets. These test results show that the suggested methodology can improve AD performance. The proposed QAE-AA architecture can be enhanced with a time series AD background and examples.

V. PROPOSED SYSTEM

In this paper, a hybrid anomaly detection technique for industrial sensor networks that combines cloud-based and edge-based sensor data analysis models is proposed. The sensor data detection model can identify unusual sensor data and upload it.
ltering enormous amounts of typical sensor data and reducing traffic load, they are sent to the cloud for additional analysis. The sensor data analysis approach can accurately identify the attack by efficiently extracting temporal and geographical information.
For sensor measurement corroboration, we first develop the spatial integrity test. The residual error-based approach makes use of sensors to measure an object's distance from a number of different angles, along with the standard measurement error to guarantee the readings are precise. Iterative, sequential training is used to train classifiers. Every iteration seeks to build upon and increase the excellence attained in the preceding iteration. At the end of the iteration, a strong classier will be created..
We evaluated the effectiveness of the approach in comparison to statistic thresholding and without the dynamic thresholding, and we discovered that it performed well in identifying anomalous data in a failure scenario while significantly reducing the time it took to identify true anomalous behaviour. The developed method can be simply used to data with diverse specifications because it is not limited to the subject of the dataset. Numerous tests are carried out with real-world data to confirm the superiority of our approach. Experimental findings demonstrate the viability of our technique in addressing the issues with conventional anomaly detection while preserving accuracy and efficacy. Dynamic thresholding and weighted loss can be implemented to a variety of deep learning architectures with ease, and they may offer similar improvements to those observed in this study.

VI. DATA FLOW DIAGRAM



VII. SYSTEM DESIGN

A. Module 1 : Data Preprocessing

Because different measurement attributes (modals) frequently have different measurement ranges, the orders of magnitude between various modes vary significantly. If the original data are used directly for analysis, the comprehensive analysis's roles for modes with higher values are enhanced, while those for modes with lower values are lessened. It is necessary to standardise the original data such that each mode has the same values in order to remove the influence of the measurement range variations between modes

B. Module 2 : Dataset from NASA: Feature Engineering

Using domain knowledge from the data, feature engineering is the process of creating features that machine learning algorithms may use. By creating features from raw data that support machine learning, feature engineering can, when done correctly, It becomes essential to properly plan features or select the most important ones. As is well known, data that is gathered for model building or prediction has a variety of properties. Because columns aren't always necessary and interdependent, we can remove some columns and characteristics to make the data less dimensional. The fast Fourier transform (FFT) is a useful tool for producing Fourier transforms.. The main advantage of an FFT is speed, which is attained by lessening the number of calculations needed to evaluate a waveform. An issue with the FFT is that it can only transform a finite set of waveform data, and to take into account spectral leakage, the waveform must be weighted using a windowing function.@@ When FFT analyzers offer frequency domain data, frequency spectra are the findings that are produced. These spectra are often retrieved as power spectra or cross power spectra.

C. Module 3 : Model Development and Prediction

In this method, the normal and defective samples are divided into a labelled training set before being sent to a classifier, which then selects the best border. This technique effectively distinguishes different flaws or inconsistencies of a known type. Any signature-based system, however, has the disadvantage of having the potential to incorrectly classify previously undetected faults or anomalies, or even mistake them for typical samples if they exist. For learning autoencoders to develop usable lower-dimensional data representations that capture the most salient latent properties, the number of neurons in hidden layers must be much less than the number of features.. The LSTM modifies the state of the cell by adding or removing information using an intelligent gate structure. The gate structure allows for the selective passing of information. tissues LSTM include forget gates f input gates i and output gates o .

VIII. FUTURE WORKS

Our future work will incorporate validation on new sensor-based datasets, features, and graph construction techniques, as well as other performance-improving guiding functions. Future research should consider more advanced sensor data anomaly detection techniques. Further investigation can also be done into how the sensor data correlation graph was made. The effectiveness of the suggested approach must also be thoroughly assessed in the presence of advanced persistent threat attacks.

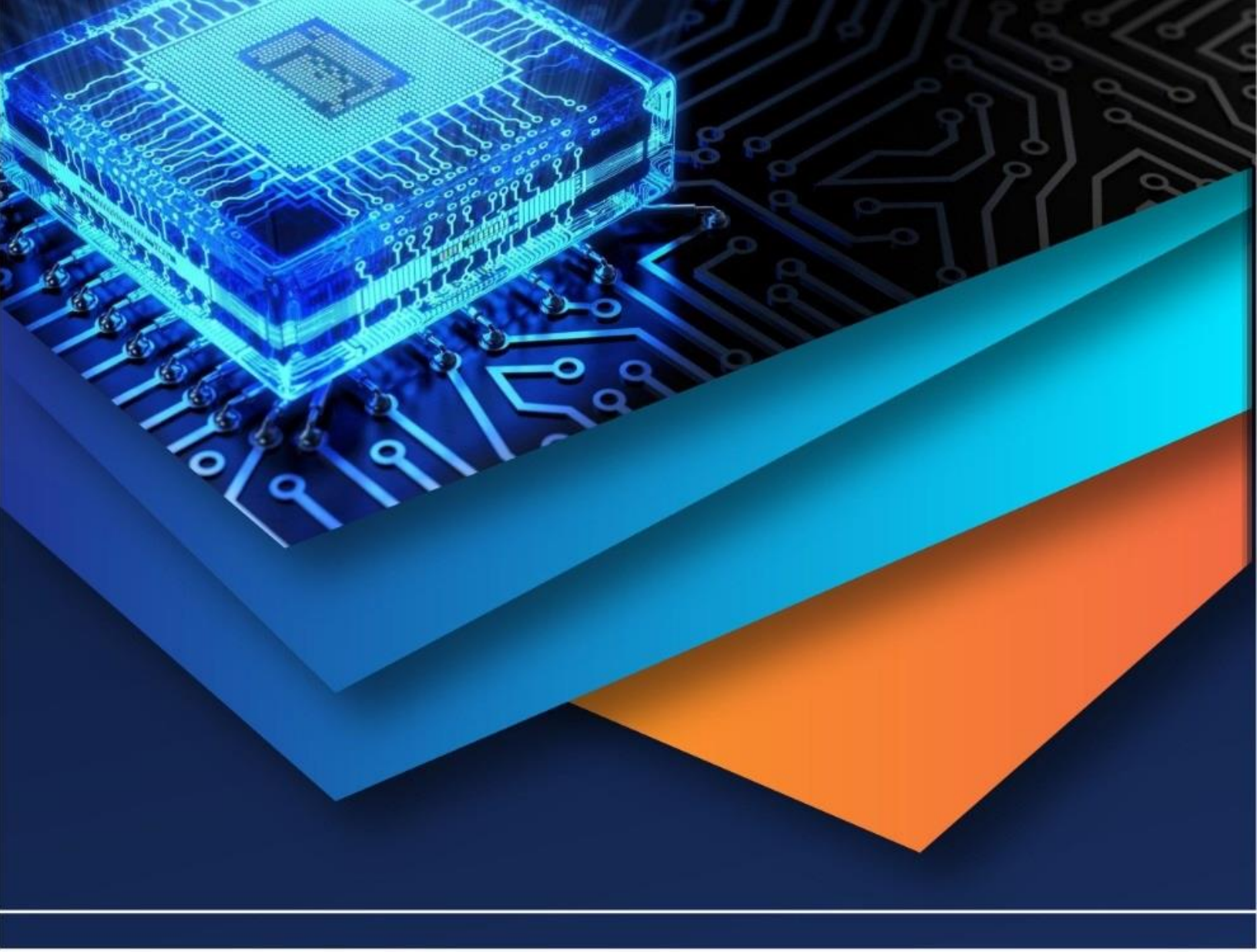
IX. CONCLUSION

This study investigated the rapid detection of anomalies in networks of unlabeled samples. It is recommended in this paper to utilise any form of neural network to fully use the temporal features and to make state assessments and to adaptively fuse the common and special properties of nodes. This framework for anomaly detection combines the advantages of supervised classification models and reconstruction models. The fundamental relationships between the sensor channels are understood using a neural network as the kernel, and abnormalities are detected using a multi-phased approach. The NN model was trained using data with normal operational conditions. The model does time series forecasting using a multi-channel sequence for the present time window and produces a prediction for the upcoming observation. The overall anomaly is determined by comparing the projected value to the actual observation. Utilising a number of targeted scenarios and tests, the system may link unusual behaviour to a specific piece of data. According to the results of the investigation, our technology successfully detects anomalies and risky situations.

REFERENCES

[1] Mining frequent trajectory patterns of WIP in an Internet of Things-based spatial-temporal database is a project by H. Cai, G. Yu, W. A. Yang, and K. Lu.
[2] Y. Zhao and F. Ferrari, Topological impacts on the mechanical characteristics of doi: /j.physa.,S.
[3] Anomaly detection through short local trajectories, Biswas and R. V. Babu, Journal of Neucomputing,.,.,., Z. Sun, D. Cao, H. He, X. Li, Grant Imahara, Mouser Electronics, Manseld, TX, USA, developed a new integrated local trajectory planning and tracking control framework for autonomous ground vehicles. Online, Internet. Offering: Moving Things. A. Gardi, R. Sabatini, and T., "Accessed: Dec.

[4] Kistan, Multiobjective, D trajectory opti- mization for integrated avionics and air traffic management systems, doi:
[5] R. Zhen, Y. Jin, Q. Hu, Z. Shao, and N. Nikitakos, "Maritime anomaly identification inside coastal waters based on vessel trajectory grouping and doi:., /s,.,
[6] Fluid-induced transition from banded kyanite- to bimineralic eclogite and implications for the development of cratons, Geochimica j.gca.,.,.,., H. Sommer, D. E. Jacob, R. A. Stern, D. Petts, D. P. Mattey, and D. G. Pearson.
[7] Pyrite multiple-sulfur isotope evidence for fast expansion and contraction of the early Paleoproterozoic seawater sulphate was published in a study by C. Scott, B. A. Wing, A. Bekker, N. J. Planavsky, P. Medvedev, S. M. Bates, M. Yun, and T. W. Lyons. ,
[8] Adjoint BFKL at nite coupling: A short-cut from the collinear limit, B. Basso, S. Caron-Huot, and A. Sever, Journal of High Energy Physics. Berlin, Germany: Springer, January, p., doi:,., /JHEP, (,),.,
[9] Bayesian detection of clusters and discontinous patterns, L. Knorr-Held and G. Raer, doi:., /j., -, X.,.,.x., The NIST definition of cloud computing, NIST, Gaithersburg, MD, USA, Tech. Rep. Special Publication, P. Mell and T. Grance, -, , . ,
[10] Cloud computing: State-of-the-art, Q. Zhang, L. Cheng, and R. Boutaba, May , . , Reliability prediction and sensitivity analysis based on software design, S. S. Gokhale and K. S. Trivedi, Proc.,
[11] Y. Maleh, A. Ezzati, Y. Qasmaoui, and M. Mbida published a paper titled "A global hybrid intru- sion detection system for wireless sensor networks" in the Procedia Computer journal.
[12] M. A. Sharkh and M. Kalil, "A quest for optimising the data processing decision for cloud-fog hybrid environments," Proc. Int. Conf. Commun., G. Thamilarasu and S. Chawla, "Towards deep-learning-driven intrusion," Apr.
[13] SafeDrive: Online driving anomaly identification from large-scale vehicle data, M. Zhang, C. Chen, T. Wo, T. Xie, M. Z. A. Bhuiyan, and X. Lin, Aug.
[14] M. Razzaq, G.-R. Kwon, and S. Shin, Energy efcient Dijkstra-based weighted sum minimization routing protocol for WSN, in Proc. , rd Int



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY