EAI Endorsed Transactions
on Internet on Things
Research Article EAI.EU

Proper Weather Forecasting Internet of Things Sensor Framework with Machine Learning

Anil V Turukmane[1*], Sagar Dhanraj Pande [2]

[1,2]School of Computer Science & Engineering, VIT-AP University, Amravati, Andra Pradesh, India

Abstract

Recent times have seen a rise in the amount of focus placed on the configurations of big data and the Internet of Things (IoT). The primary focus of the researchers was the development of big data analytics solutions based on machine learning. Machine learning is becoming more prevalent in this sector because of its ability to unearth hidden traits and patterns, even within exceedingly complicated datasets. This is one reason why this is the case. For the purpose of this study, we applied our Big Data and Internet of Things (IoT)-based system to a use case that involved the processing of weather information. We put climate clustering and sensor identification algorithms into practice by using data that was available to the general public. For this particular application, the execution information was shown as follows:every single level of the construction. The training method that we've decided to use for the package is a k-means cluster that's based on Scikit-Learn. According to the results of the information analyses, our strategy has the potential to be utilized in usefully retrieving information from a database that is rather complicated.

Keywords: Internet of Things; Big data; Complex dataset; Machine learning; Clustering

Received on 18 December 2023, accepted on 07 March 2024, published on 12 March 2024

doi: 10.4108/eetiot.5382

*Corresponding author. Email: anilturukmane@gmail.com

1. Introduction

The term "short-range weather reports" refers to forecasts that cover a period of one to two days in the future. On these timescales, one would be interested in the rate at which a cold front arrives or the progression and spread of particular thunderstorm formations. Because cloud platforms require very elevated simulations with horizontally spaced geographical units of the scale of 1 km [1-3], this is a difficult problem to solve. Because of this, short-range estimate techniques will invariably have a regional focus rather than a global approach when they are applied to processing capabilities. This is an unavoidable consequence. Even when employing a 1 km lattice, such turbulent cloud formations frequently include terms in updrafts that are too tiny to be fully handled, and their expansion is strongly dependent on the deposition of precipitation [4]. Modeling on these time and space scales is therefore still problematic. The greatest significant improvements in ability over the past few decades have probably been in medium-range projection, over periods of a few days to several weeks. These projections cover periods of time ranging from a few days to several weeks. The baroclinic waves are the prototypical event that occurs on the medium-range timeframe, and numerical weather forecasting systems that are globally connected and have grid time intervals that are on the order of ten kilometers are excellent at resolving the distinctive features that baroclinic waves exhibit [5]. In many cases, it is possible to arrive at an accurate evaluation of these lengths spanning a span of 20 years of baroclinic waves. Subseasonal forecasts can cover time periods ranging from two weeks to a whole season. On this scale, the Madden Julian Oscillation is a significant factor in forecasting. This equatorially limited oscillation in tropical winds has a period of 30 to 80 days, and its propagation is to the east. According to [6-7], this

oscillation is associated to either increased or inhibited structured precipitation. In recent times, there has been a lot of interest directed toward this timeframe as a result of the commencement of the Subseasonal to Seasonal Prediction Experiment. The linked processes of the atmosphere and the ocean are essential for annual forecasting, and the occurrence of El Nino in the equatorial region of the Pacific Ocean is a classic example [9]. Many large datasets connected to this program include S2S projections from tactical and academic simulations, which are useful for possible ML requests. Within the context of cyclical timescales, coupled ocean-The atmosphere and its systems have shown that they are capable of El Nino forecasting. When simulating at this frequency, however, it is necessary to use grid distances that span several tens of kilometers.

In addition, climate change forecasting investigates how meteorological data change over long periods of time—years and even decades—as a result of an increase in the concentration of greenhouse gases in the atmosphere [10]. Similarly, for these durations, the simulation approaches reveal very strong biases whose amplitudes are equivalent to the expected climate change indication [11]. Again, this is the case for all of the simulation methods. It would appear that making proposals for now-casting would be an appropriate area for Tough AI to get started. Because errors won't add up to much after a few days, technical limitations such as conservation laws can be disregarded on these durations. In addition, the abundance of information provided by the Internet of Things (IoT) is becoming increasingly easily accessible and is being used for weather forecasting in a variety of formats, such as the information provided by cell phones [12]. The integration of this information using conventional methods is going to be fairly difficult because there are a great number of observations and considerable errors.

2 Literature Survey

The term "short-range weather reports" refers to forecasts that cover a period of one to two days in the future. On these timescales, one would be interested in the rate at which a cold front arrives or the progression and spread of particular thunderstorm formations. Because cloud platforms require very elevated simulations with horizontally spaced geographical units of the scale of 1 km [1-3], this is a difficult problem to solve. Because of this, short-range estimate techniques will invariably have a regional focus rather than a global approach when they are applied to processing capabilities. This is an unavoidable consequence. Even when employing a 1 km lattice, such turbulent cloud formations frequently include terms in updrafts that are too tiny to be fully handled, and their expansion is strongly dependent on the deposition of precipitation [4]. Modeling on these time and space scales is therefore still problematic. The greatest significant improvements in ability over the past few decades have probably been in medium-range projection, over periods of a few days to several weeks. These projections cover periods of time ranging from a few days to several weeks. The baroclinic waves are the prototypical event that occurs on the medium-range timeframe, and numerical weather forecasting systems that are globally connected and have grid time intervals that are on the order of ten kilometers are excellent at resolving the distinctive features that baroclinic waves exhibit [5]. In many cases, it is possible to arrive at an accurate evaluation of these lengths spanning a span of 20 years of baroclinic waves. Subseasonal forecasts can cover time periods ranging from two weeks to a whole season. On this scale, the Madden Julian Oscillation is a significant factor in forecasting. This equatorially limited oscillation in tropical winds has a period of 30 to 80 days, and its propagation is to the east. According to [6-7], this oscillation is associated to either increased or inhibited structured precipitation. In recent times, there has been a lot of interest directed toward this timeframe as a result of the commencement of the Subseasonal to Seasonal Prediction Experiment. The linked processes of the atmosphere and the ocean are essential for annual forecasting, and the occurrence of El Nino in the equatorial region of the Pacific Ocean is a classic example [9]. Many large datasets connected to this program include S2S projections from tactical and academic simulations, which are useful for possible ML requests. Within the context of cyclical timescales, coupled ocean- The atmosphere and its systems have shown that they are capable of El Nino forecasting. When simulating at this frequency, however, it is necessary to use grid distances that span several tens of kilometers.

In addition, climate change forecasting investigates how meteorological data change over long periods of time—years and even decades—as a result of an increase in the concentration of greenhouse gases in the atmosphere [10]. Similarly, for these durations, the simulation approaches reveal very strong biases whose amplitudes are equivalent to the expected climate change indication [11]. Again, this is the case for all of the simulation methods. It would appear that making proposals for now-casting would be an appropriate area for Tough AI to get started. Because errors won't add up to much after a few days, technical limitations such as conservation laws can be disregarded on these durations. In addition, the abundance of information provided by the Internet of Things (IoT) is becoming increasingly easily accessible and is being used for weather forecasting in a variety of formats, such as the information provided by cell phones [12]. The integration of this information using conventional methods is going to be fairly difficult because there are a great number of observations and considerable errors.

Ghosh et al.'s 2023 study [13] focuses on "Water Quality Assessment Through Predictive Machine Learning", highlighting the use of machine learning for analyzing and predicting water quality parameters. In "Unraveling the Heterogeneity of Lower-Grade [14] Gliomas," Rahat, Ghosh, and colleagues (2023) delve into deep learning-assisted segmentation and genomic analysis of brain MR images, offering new insights into this medical condition. Potato Leaf Disease [15] Recognition and Prediction using Convolutional Neural Networks," by Ghosh, Rahat, and team (2023), showcases the application of convolutional neural networks in accurately identifying diseases in potato leaves. Mandava, Vinta, Ghosh, and Rahat's [16]2023 research presents "An All-Inclusive Machine Learning and Deep Learning Method for Forecasting Cardiovascular Disease in Bangladeshi Population", integrating advanced AI techniques for health predictions. The 2023 study by Mandava et al., titled "Identification and Categorization of Yellow [17] Rust Infection in Wheat through Deep Learning Techniques", applies deep learning methods to detect and categorize wheat infections effectively. Khasim, Rahat, Ghosh, and colleagues' 2023 article, "Using Deep [18] Learning and Machine Learning: Real-Time Discernment and Diagnostics of Rice-Leaf Diseases in Bangladesh", explores AI-based solutions for diagnosing rice-leaf diseases. Deciphering Microorganisms through Intelligent Image Recognition", authored by Khasim, Ghosh, Rahat, and others in 2023, discusses [19] the use of machine learning and deep learning in identifying microorganisms through advanced image recognition techniques. The 2023 study by Mohanty, Ghosh, Rahat [20] and Reddy, "Advanced Deep Learning Models for Corn Leaf Disease Classification", focuses on the application of deep learning in classifying diseases in corn leaves based on a field study. Alenezi and team's 2021 research, "Block-Greedy and CNN Based Underwater Image Dehazing [21] for Novel Depth Estimation and Optimal Ambient Light", investigates novel CNN-based methods for enhancing underwater image clarity and depth estimation.

3. Proposed System

In order to conduct an analysis of the acquired meteorological data, this study makes use of the following methodologies. Technology for Linked Observations and Sensors, Which Are Also Connected At the level of data processing, information is obtained through a procedure known as bulk uploading. After the information has been acquired, it is instantly transferred to the ETL level in preparation for the upcoming procedures. In the ETL level represented in Figure 1, sensor information and observational statistical measures are processed using regular expressions. The clustering-specific Python script is invoked by the Node server after the file has been created. The Python-shell is what's used to execute Python scripts. As production continues, the stage that we are now at when creating the Dataset will change. Instead of being formatted as a CSV, the data item that contains the results of the processing will be delivered directly to the application.

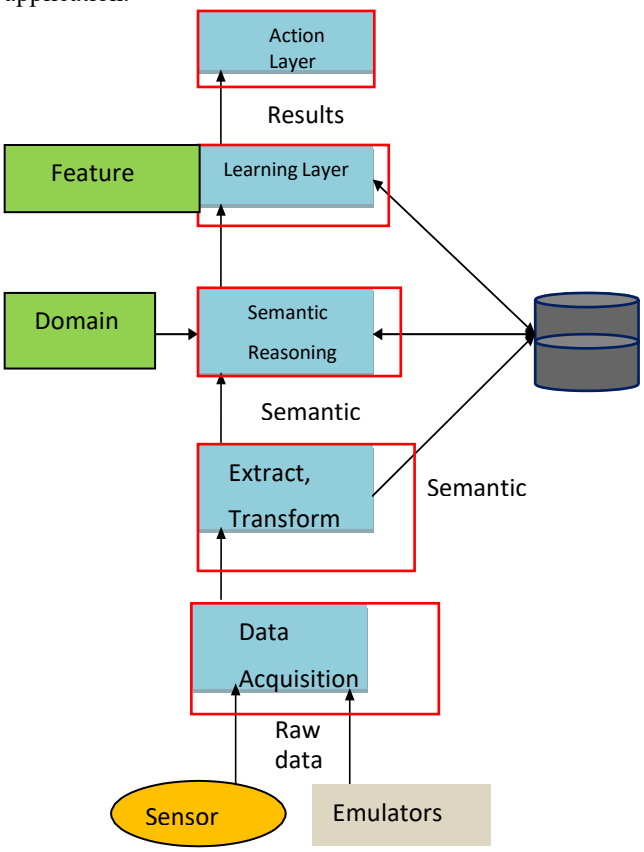

Figure 1. Architecture of proposed system

Figure 2 illustrates the many steps of the information collection process, including ETL and semantic processing, as well as training. Two distinct n3 documents called Linked Sensor Technology and Linked Data Collected are included in each and every data set. Both of these documents are present. In addition, every piece of sensor data possesses its own singular collection of characteristics. Connected Sensor Information is utilized in order to ascertain the position of sensors, in addition to their height, longitude, and meridian. With linked observation data, one may determine the altitude, humidity levels, ambient temperature, relative humidity, prevailing winds, wind burst, wind direction, and transparency [23]. Because Linked Observation Data contains a large amount of information that isn't necessary, clustering techniques are utilized in order to extract the appropriate fraction of information from the document. The information from the associated sensors is analyzed using the data structure, and the results are saved.the moving of the

marker in the activity-specific map view. In order to make room for all of the sensor properties and recorded location services, a subclass structure must be created for each and every type of sensor. After that, related observation data will be analyzed in the subsequent step. Once again making use of sequence, the observation variables are retrieved, and this time they are saved in the previously stated class example.

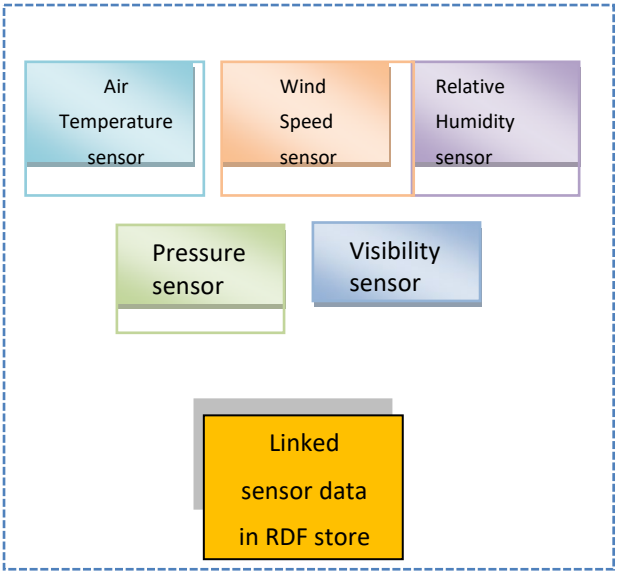

Figure 2 illustrates the many steps of the information collection Process

4. Existing System

India's economic foundation rests on the success of its agricultural sector. However agricultural land and significance are declining for a variety of reasons. Efforts to improve farming practices and preserve farmland are the subject of a great deal of study and investigation. For identifying soil/seed/crop quality, weed detection, predicting

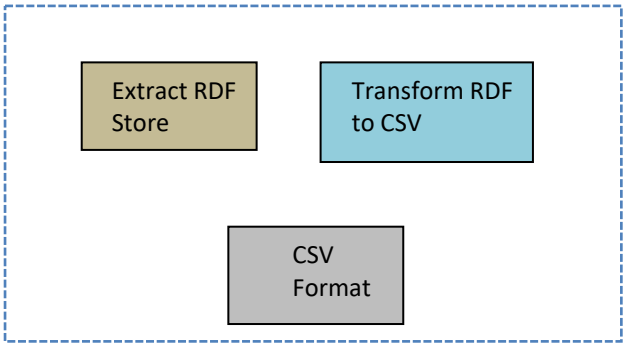

Fig 3: Use-case Scenario framework

The not existent features abbreviation, N/A, is used because the extracted features may be present in varying proportions in each individual set of sensor data. In a nutshell, a translation script is developed with the goal of being able to construct a dataset and provide sensor results for the specified time period, such as 6:00 PM to 7:00 PM. It is not difficult to do.

With this RDF format, obtain the sensor identity data along with the elevation, longitude, and meridian coordinates.

In this particular instance, the characteristic description that was gathered was called MeasureData, and it was on August 23, 2022, at 5:20 PM. It received a rating of 300. Due to the fact that all of these additional qualities utilize the same syntax, retrieving them is made far simpler by the fact that we parsed the data using Regex. During the learning phase, it is possible to obtain previously hidden data, and methods from the field of machine learning are utilized in order to collect specific data. Processing the data from the weather sensors is necessary in this scenario in order to gain more information. Sensors are unable to identify the information they collect.

So, we will group the dataset in order to search for any hidden data. The K-Means clustering methodology was used in our research to find broad patterns in the information as well as sensor defects and anomalies. This was accomplished by grouping similar sensor data together. The formation of strategies and the use of learning algorithms are both in progress. The first algorithm presents a recommended and generalized form of supervised learning.

5. Expected Result

We were able to create a data clustering by making use of the aforementioned publicly available and freely accessible weather information. We only included groups with 2, 3, and 4 people in order to make the findings easier to read. Because the areas were disappearing when there were more than 4 groups involved, we decided that 4 should be the largest possible value for k. In the following subcategories, greater explanation is provided for the particular feature options that are available for the various grouping possibilities.

The Himalaya Mountains unquestionably perform an outstanding function as an effective barrier between the two distinct meteorological information groups, and the information grouping outputs reveal a reasonable separation between geographical areas of south Asia. The fact that the model produces group chunks that are almost the same size as one another is something that is very

interesting. As can be seen in Figure 3(a), the information is more or less evenly dispersed over the world. In contrast to the southern portion of the Himalaya Mountains, the region down the coast of those mountain ranges experiences temperatures that are a few degrees lower and a major reduction in the amount of precipitation that falls each year. The levels of humidity tend to be the most distinguishing feature; as a consequence, even if the temperature is typically 8 degrees higher in the Western region, the moisture content is significantly lower, which may help to balance the higher degree. The reality is It is essential to take into account the fact that the temperature photograph was captured on August 29, which was one of the hottest days of the entire year.

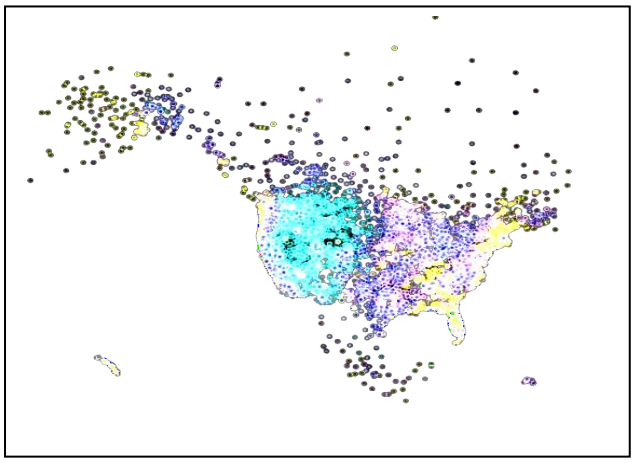

Fig 4 : AirTemp, Relative Humidity, WindSpeed 2 Clusters

When utilizing 2 groups, the spatial variation becomes immediately apparent. When we do three groupings, as shown in Figure 3(b), we also encounter fascinating outcomes, such as the emergence of connections between east- and west-coast sections. However, the mountainous plains that are located between any Mountains and the western coast continue to stand out as their own unique zone. The Tibetan Plateau and the Deserts both function as transitional areas, in contrast to the Mountains and the Eastern Coast, which are both fixed features. There is a transferrable zone created by the Asia and Southeast Himalaya Mountains that is comparable to the Eastern Coast, and we also notice common traits on the southern Coast. On this particular occasion, the transition zone has temperature and humidity values that are precisely in the middle of the aforementioned two regions, and the differences between the degrees and the levels of humidity reflect the exact same principles that are at play when two clusters are present. Once more, it would appear that the speed of the wind is not a significant component when combined with the other variables that affect the climate. It is essential to make the observation that, similar to the situation with the two groups, the distribution of the data among all of the different categories is, for the most part, consistent.

Even though it seems to make sense that the relevant geographical locations for this evaluation would have been the same as those for the analysis of ambient temperature, moisture, and wind direction, there are a few notable differences between the two sets of outcomes. These differences were found despite the fact that it seems to make sense that the relevant geographical locations for this evaluation would have been the same as those for the analysis. To begin, in contrast to the prior circumstance, the regions are not divided evenly into distinct halves. An analysis of three clusters uncovers some fascinating patterns, as demonstrated in Figure 4(b). In spite of the fact that the findings are comparable to those of the 2-cluster example, the group that emerged in the region between the Mountain Range and the east Side spread all the way to the Plain and the Munak Canal (Figure 4), as shown (a). The Northern Valleys, North and West Valleys, South Eastern Valleys, and Southern Valleys all create larger temperature zones around the Northern Plains, which indicate a more compact temperature center. This fits in reasonably well with the seasonal weather projections for the summer. This region of higher temperatures frequently goes in a diagonal direction from Southern Asia to Telegana and then further north to the Waterways. An analogous horizontal structure can be observed all the way from the Mahanadi River valley in the south to the Arabian Sea in the north.

6. Conclusion

This study provides a description of an improved Internet of Things (IoT) system with a use instance on a meteorological information clustering technique that involves the retrieval of information, various levels of analysis, and learning. We developed a learning strategy that makes use of the grouping unsupervised learning strategy during the training process of the strategy so that we may make the most of the enormous datasets that are associated with this subject matter. In India the weather information that is collected from India's 8000 different weather stations across the country is accessed through the use of event logs. This information is compiled and analyzed with the help of calls made in Node.js. There were three different wind speed groups, a device failure category, and a training period category. The data processing for this particular study makes use of information regarding the ambient temperature, the direction of the wind, the humidity levels, the transparency, and the altitude. The well-known k-means clustering algorithm is put to use, and the obtained results are displayed. An unusual feature was brought to our attention when we discovered that the

clustering technique reflected the spatial coherence of the stations. To put it another way, various key geographical sites on the Indian continent each have their own distinct weather group, which can be differentiated from the others in a straightforward manner. In addition, the process of clustering identifies any potential sensor faults or anomalies that may exist. We were able to demonstrate how Internet of Things Big Data architecture could perhaps be utilized in such deployments with the assistance of this use case.

References

[1] Chelliah, B. J., Latchoumi, T. P., & Senthilselvi, A. (2022). Analysis of demand forecasting of agriculture using machine learning algorithm. Environment, Development and Sustainability, 1-17.

[2] Kruse, J., Schäfer, B., & Witthaut, D. (2021). Revealing drivers and risks for power grid frequency stability with explainable AI. *Patterns*, *2*(11), 100365.

[3] Cho, D., Yoo, C., Son, B., Im, J., Yoon, D., & Cha, D. H. (2022). A novel ensemble learning for post-processing of NWP Model's next-day maximum air temperature forecast in summer using deep learning and statistical approaches. *Weather and Climate Extremes*, *35*, 100410.

[4] Latchoumi, T. P., Swathi, R., Vidyasri, P., & Balamurugan, K. (2022, March). Develop New Algorithm To Improve Safety On WMSN In Health Disease Monitoring. In *2022 International Mobile and Embedded Technology Conference (MECON)* (pp. 357-362). IEEE.

[5] Chattopadhyay, A., Mustafa, M., Hassanzadeh, P., Bach, E., & Kashinath, K. (2021). Towards physically consistent data-driven weather forecasting: Integrating data assimilation with equivariance-preserving deep spatial transformers. *arXiv preprint arXiv:2103.09360*.

[6] Duan, Z., Liu, H., Li, Y., & Nikitas, N. (2022). Time-variant post-processing method for long-term numerical wind speed forecasts based on multi-region recurrent graphnetwork. *Energy*, *259*, 125021.

[7] Betancourt, C., Stomberg, T., Roscher, R., Schultz, M. G., & Stadtler, S. (2021). AQ-Bench: a benchmark dataset for machine learning on global air quality metrics. *Earth System Science Data*, *13*(6), 3013-3033.

[8] Fernández, J. G., Abdellaoui, I. A., & Mehrkanoon, S. (2022). Deep coastal sea elements forecasting using UNet-based models. *Knowledge-Based Systems*, *252*, 109445.

[9] Niu, D., Huang, J., Zang, Z., Xu, L., Che, H., & Tang, Y. (2021). Two-Stage Spatiotemporal Context Refinement Network for Precipitation Nowcasting. *Remote Sensing*, *13*(21), 4285.

[10] Diaconu, C. A., Saha, S., Günnemann, S., & Zhu, X. X. (2022). Understanding the Role of Weather Data for Earth Surface Forecasting Using a ConvLSTM-Based Model. In *Proceedingsof the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 1362-1371).

[11] Frezat, H., Sommer, J. L., Fablet, R., Balarac, G., & Lguensat, R. (2022). A posteriori learning for quasi-geostrophic turbulence parametrization. *arXiv preprint arXiv:2204.03911*.

[12] Balamurugan, K., Latchoumi, T. P., & Ezhilarasi, T. P. (2022). Wearables to Improve Efficiency, Productivity, and Safety of Operations. In *Smart Manufacturing Technologies for Industry 4.0* (pp. 75-90). CRC Press

[13] Ghosh, H., Tusher, M.A., Rahat, I.S., Khasim, S., Mohanty, S.N. (2023). Water Quality Assessment Through Predictive Machine Learning. In: Intelligent Computing and Networking. IC-ICN 2023. Lecture Notes in Networks and Systems, vol 699. Springer, Singapore. https://doi.org/10.1007/978-981-99-3177-4_6

[14] Rahat IS, Ghosh H, Shaik K, Khasim S, Rajaram G. Unraveling the Heterogeneity of Lower-Grade Gliomas: Deep Learning-Assisted Flair Segmentation and Genomic Analysis of Brain MR Images. EAI Endorsed Trans Perv Health Tech [Internet]. 2023 Sep. 29 [cited 2023 Oct. 2];9. https://doi.org/10.4108/eetpht.9.4016

[15] Ghosh H, Rahat IS, Shaik K, Khasim S, Yesubabu M. Potato Leaf Disease Recognition and Prediction using Convolutional Neural Networks. EAI Endorsed Scal Inf Syst [Internet]. 2023 Sep. 21 https://doi.org/10.4108/eetsis.3937

[16] Mandava, S. R. Vinta, H. Ghosh, and I. S. Rahat, "An All-Inclusive Machine Learning and Deep Learning Method for Forecasting Cardiovascular Disease in Bangladeshi Population", EAI Endorsed Trans Perv Health Tech, vol. 9, Oct. 2023.https://doi.org/10.4108/eetpht.9.4052

[17] Mandava, M.; Vinta, S. R.; Ghosh, H.; Rahat, I. S. Identification and Categorization of Yellow Rust Infection in Wheat through Deep Learning Techniques. EAI Endorsed Trans IoT 2023, 10. https://doi.org/10.4108/eetiot.4603

[18] Khasim, I. S. Rahat, H. Ghosh, K. Shaik, and S. K. Panda, "Using Deep Learning and Machine Learning: Real-Time Discernment and Diagnostics of Rice-Leaf Diseases in Bangladesh", EAI Endorsed Trans IoT, vol. 10, Dec. 2023 https://doi.org/10.4108/eetiot.4579

[19] Khasim, H. Ghosh, I. S. Rahat, K. Shaik, and M. Yesubabu, "Deciphering Microorganisms through Intelligent Image Recognition: Machine Learning and Deep Learning Approaches, Challenges, and Advancements", EAI Endorsed Trans IoT, vol. 10, Nov. 2023. https://doi.org/10.4108/eetiot.4484

[20] Mohanty, S.N.; Ghosh, H.; Rahat, I.S.; Reddy, C.V.R. Advanced Deep Learning Models for Corn Leaf Disease Classification: A Field Study in Bangladesh. Eng. Proc. 2023, 59, 69. https://doi.org/10.3390/engproc2023059069

[21] Alenezi, F.; Armghan, A.; Mohanty, S.N.; Jhaveri, R.H.; Tiwari, P. Block-Greedy and CNN Based Underwater Image Dehazing for Novel Depth Estimation and Optimal Ambient Light. Water 2021, 13, 3470. https://doi.org/10.3390/w13233470