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

Iot Techniques for Disaster Prediction and Prevention

[1]Mustafa Hadi Abdullah, [2]Zaid Hamodat

Submitted: 23/04/2023 **Revised:** 25/06/2023 **Accepted:** 05/07/2023

Abstract: Natural catastrophes such as landslides, floods, fires, and volcanic eruptions, as well as the damage produced by these events, are global issues that result in financial and human losses. This problem is exacerbated by changes in the planet's environmental conditions and is primarily evident in metropolitan areas. Because of pollution and a lack of planning, the deterioration of the ecosystem is more pronounced in these areas, damaging the ecology and influencing the local climate. As a result, this initiative makes three major contributions: (i) the use and evaluation of new IoT standards and emerging technologies in conjunction with WSN for the collection and distribution of data in natural environments, (ii) the use of the collected data for the prediction of natural disasters using Machine Learning (ML) techniques, with a case study on the characteristics of rivers and rainfall in Iraq and Turkey, and (iii) the proposal of an IoT-based and ML-based fault-tolerant architecture for the system.

Keywords*: ML, DL, WSN, IOT.*

I. Introduction

For example, the frequency and magnitude of floods. One example of these changes is the formation of heat islands, which end up altering the region's rainfall pattern. It is claimed that global warming is the primary cause of the increase in the number of natural disasters and that climate monitoring via sensors is something viable and necessary for making timely predictions and alerts. for making decisions about earthquakes, landslides, and floods, among other things. In the case of floods,) state that over 102 million people are impacted by floods each year around the world, and this figure is anticipated to rise in the coming years, as the regions most affected by floods are emerging countries and urban areas. These are precisely the features of Turkey, where the climate has altered in recent years, making the environment more conducive to natural disasters; second, they repeat and add that present solutions do not greatly reduce flood damage. In actuality, many natural disasters, such as floods, cannot be prevented, but their impacts can be lessened and controlled if they are anticipated ahead of time.

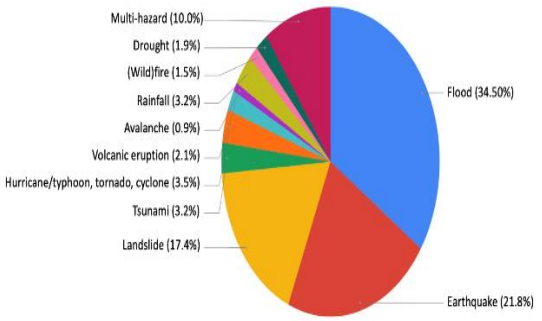


Fig 1 Probability of the types of disasters

So far, the system can only identify floods after they have already begun. However, the system should predict floods and send notifications in time for the populace to take preventive actions and evacuate high-risk locations. This can help to alleviate a number of flood-related issues. However, due to the large number of proprietary and non-proprietary solutions, the high heterogeneity in the technologies used by the Wireless Sensor Network (WSN) already in use impedes the development and integration of WSN to new technologies and the WSN already in use.

In a natural disaster predicting and management scenario, the sensors' data can be examined in conjunction with additional information available via the Internet, such as satellite photos and forecasts, or variable values to which the sensors would not have access. Furthermore, if the sensors are connected to the Internet, technology such as cloud computing and social networks can be used to create predictions and generate online warnings about the monitored settings.

1Electric and computer engineering Altinbas University
Istanbul, Turkey
203720112@ogr.altinbas.edu.tr
2Electric and computer engineering Altinbas University
Istanbul, Turkey
zaidghanim88@gmail.com

The IoT vision allows us to extend the REDE system and add the concept of device or remote service accessibility in the context of this project, which aids in the dissemination of knowledge, decision making, or action in the real world through sensing and information generation. As a result, WSN nodes (such as the REDE system) can be deemed smart objects if they meet specific criteria, such as data gathering, data utilization over the Internet, and collective intelligence among all nodes, rather than just the situation and data of a single node. In this respect, we might claim that natural disaster management and prediction are scenarios in which organic integration with the real world happens, and therefore the use of WSN for this is justified.

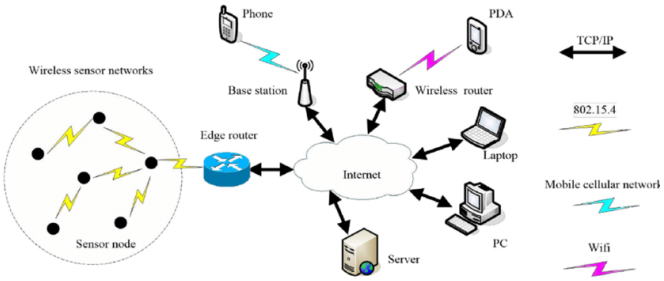


Fig 2 WSN-IOT sensing system

In this approach, we might argue that management and forecasting The system's next phase in growth will be to predict floods and expand to other sorts of natural calamities. These predictions must be carried out even if a component of the system fails, because the system is designed to operate in hostile situations and hence requires fault tolerance techniques. Another promising avenue for growth is the integration of the REDE system with upcoming technologies such as IoT standards, as one of the primary goals of this effort is to investigate the aforementioned areas. Natural disasters are circumstances in which spontaneous integration with the real world occurs, and the usage of WSN for this purpose becomes necessary.

The main problem of this paper is related to the damage caused by natural disasters, mainly in urban areas, and the lack of mechanisms that can predict these disasters and mitigate the damage caused. It is necessary to use data collected by systems similar to REDE to predict disasters and generate alerts for the population in risky regions. In addition, new emerging technologies, such as IoT standards and technologies, can be used together with WSN to facilitate the acquisition, distribution, and processing of collected data and increase the accuracy of predictions. Since natural disaster forecasting systems are normally located in areas of risk and must function mainly during the occurrence of a disaster, Fault tolerance mechanisms are highly desired. Therefore, the implementation of fault tolerance mechanisms in these systems is another key point of this work.

Many works have been carried out in this area, but it is still necessary to increase the accuracy of forecasts, especially in extreme situations in which part of the system is compromised. This need is even more apparent when the focus is on the combination of IoT, WSN, ML techniques and fault tolerance mechanisms in order to make such predictions. Thus, the hypopaper of this paper is that the combination of WSN with IoT technologies and ML techniques can generate systems of great accuracy, reliability and availability and that, even in times of failure of part of the system, it can maintain forecasts and alerts for the population.

Therefore, the main objective of this work is to present an approach that improves the accuracy of natural disaster forecasts and the reliability and availability of such forecasts. In order to predict such disasters, ML techniques are used, whose generated models are embedded in the system nodes, which makes it possible for these nodes to be autonomous when making predictions and to reorganize themselves, even in situations where there are failures in communication or on the system hardware.

Therefore, the specific objectives of this paper are:

(i) to evaluate the use of ML techniques, on data collected by WSN, to predict the behavior of natural environments, such as rivers;
(ii) (Analyze the combination of sensors, positioned in geographically different regions, and their data, to increase the accuracy of these predictions and distribute the processing load necessary to make these predictions
(iii) combine and analyze IoT and WSN technologies for monitoring environments prone to natural disasters;
(iv) identify fault tolerance techniques that can be used to forecast natural disasters, increasing the reliability and availability of these forecasts;
(v) evaluate the system for monitoring and forecasting natural disasters, being proposed based on the knowledge acquired in the previous items. increasing the reliability and availability of these forecasts
(vi) evaluate the system for monitoring and forecasting natural disasters, being proposed based on the knowledge acquired in the previous items. increasing the reliability and availability of these forecasts;
(vii) evaluate the system for monitoring and forecasting natural disasters, being proposed based on the knowledge acquired in the previous items.

The main justification of this project is that computational methods aimed at forecasting are increasingly used in several areas that primarily require human knowledge, such as: bank default, construction of intelligent environments, the useful life of electric generators, reliability and the maintenance needs of industrial systems and new computational problems, such as the allocation of resources in computational grids The use of WSN to predict natural disasters is another point that justifies research in the area and is very present in the literature. As an example, we have the WSN used in this project and presented where a WSN combined with robust linear regression is proposed to carry out the prediction of floods. Furthermore, sensor network plays a crucial role in IoT architecture, thus, sensor network architecture directly affects IoT architecture and, consequently, the system and functions for which it was developed. In addition, the fact that IoT is a natural evolution of WSN also justifies a study on the use of WSN and IoT in order to achieve a fault-tolerant architecture for the prediction of natural disasters. Another factor that must be addressed is the great complexity in the analysis of data collected in natural environments, which often requires human intervention for treatment. This intervention ends up generating imprecision and unreliability in the results. Nonetheless, Humans have a great ability to make generalizations and find patterns and, in computing, this ability can be simulated to make predictions of natural disasters more accurately and with greater reliability and availability. Such simulation takes place through ML techniques used together with data collected by WSN and IoT technologies.

To achieve the objective of this paper, a fault-tolerant system called System for dEtecting and forecasting Natural Disasters based on IoT (our methodology) was developed. This system is based on WSN, IoT and ML and still provides flexibility to adapt to different environments and situations, while fault tolerance mechanisms ensure that predictions will be carried out, even with different levels of accuracy. our methodology system was modeled and tested using MATLAB, a case study on floods was carried out using the data collected by the system on the model. More details about the methodological procedures, the performance of the modeling and the analysis of our methodology are presented in Chapter4. In order to investigate the mechanisms used to predict natural disasters, even in inhospitable environments, based on technologies and patterns found in the literature. The influence of the pair in each collection point of the system is another justification for this work. In other words, the way that the reading of a sensor x, located in a region r0, influences the reading of a sensor y, located in a region r ahead of r0, considering the flow of the river. In addition, the time for this influence to occur and how technologies aimed at IoT can help in the distribution of data, in the standardization of the means of communication and in the availability of the processing power and storage capacity necessary for the manipulation of the data. present in the literature that use or not the technologies present in our methodology, a systematic review was carried out according to the protocols described in the work. such protocols must be followed in order to ensure a complete and reproducible review. the systematic review indicates that, despite being a much-explored area, solutions that aim to predict natural disasters and that make use of technologies and the combined characteristics of our methodology system still need to be studied. thus, our methodology seeks to contribute with a fault-tolerant approach that enables the prediction of natural disasters.

II. Internet of Things

The Internet of Things (IoT) is a technological revolution that evolved from prior technologies such as wireless sensor networks (WSN) and mobile ad hoc networks (MANET) [1]. The sensor is the most important component of these networks. A sensor is a small device housed in a personal area network (PAN) that has detection, measurement, computation, and communication features that allow for the observation and reaction to events.

A WSN is made up of densely dispersed sensor nodes that support sensing, connection, and signal processing at a specific place. They are linked and self-organized. The sensors send data to monitoring nodes known as Sink, which route the collected data to a base station for further analysis. Wireless sensor networks feature unique properties such as minimal battery power, a limited transmission range, and a short duty cycle. They could communicate point-to-point or multipoint-to-point (sensor nodes sending data to a central node, Sink). [2].

A mobile sensor network (MANET) is a network of mobile sensors (nodes) that communicate with one another without the requirement of infrastructure such as base stations. Because of their self-configurability and ease of deployment, these networks are known as self-organizing networks (SON) [3]. Nodes collaborate to offer connectivity and operate without centralized infrastructure or management; thus they communicate with directly linked neighbors node-to-node (peer-to-peer). Their transmission power and bandwidth are limited because to their limited mobility and batteries. Mobile nodes frequently work together to send data and route information to nodes to whom they are not physically connected,

It should be emphasized that the advantages of MANETs and WSNs were swiftly recognized from the beginning

of their development, which led to their deployment in various domains of application such as agriculture, industry, and systems. health. Today, the Internet of Things has emerged from the ashes of these two technological cornerstones.

The Internet of Things (IoT) is the responsive usage of systems, heterogeneous technologies, and the developing paradigm of device interconnectivity across physical settings via TCP/IP [4]. From the outside, the Internet of Things appears to be an M2M communication (Machine-to-Machine: a communication between system entities connected with or without wire that does not necessarily require direct human intervention), but the IoT includes billions of connected objects such as humans, appliances, vehicles, machines, pets, cattle in the field, animals in the wild, habitats, occupants of housing, and even businesses and their interactions. IoT has become a fad today and promises to change the way the internet is viewed and used. In a McKinsey Global Institute survey report, the impact of IoT on the global market will be $11 trillion by 2025 [5]. Likewise, the Internet of Things is not just a technology of the future, it is here with us. The Internet of Things includes the new wave of sensors and works with the growing cloud network infrastructure [6]. isn't just a technology of the future, it's here with us. The Internet of Things includes the new wave of sensors and works with the growing cloud network infrastructure [6]. isn't just a technology of the future, it's here with us. The Internet of Things includes the new wave of sensors and works with the growing cloud network infrastructure [6].

A. IoT Architecture

It is critical to have a standard model of IoT architecture since it provides standards, implementations, and viewpoints for developing interoperable IoT systems. There is currently no generalized IoT architecture. Several IoT architectural standards, however, have been proposed by various scholars and research organizations. Some of the most common IoT topologies available today are explored here.

The authors of [18] presented a hierarchical structure model with three layers: perception layer, network layer, and application layer. The sensory organ of the IoT is represented by the perception layer, which is located at the bottom. Its goal is to recognize objects and collect data. RFID tags, 2D barcode scanners, terminals, GPS units, cameras, sensors, and sensor networks are all part of this layer. The network layer is the second layer. This layer is the heart of the Internet of Things. It processes and transmits data from the perception layer to the application layer. The network layer consists of the following components: the information center, the intelligent processing center, the internet network systems, and the network management center.

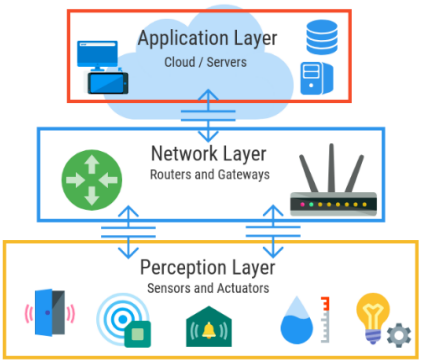


Fig 3 Three-layer architecture model [18].

A second open, public, and royalty-free architecture is given by FI-WARE, a European Union-funded initiative [19]. This architecture is built with open components known as Generic Enablers (GE). These GEs provide reusable and shared functionality. GEs are classified into six major groups, each of which provides an architectural reference model for the specific features addressed in the architecture:

- Cloud Hosting: This involves compute, storage, and network resources.
- Data management: represents the access, processing and analysis of data flows, transforming them into a database available to applications.
- Application Ecosystem Framework: This is the infrastructure for building, distributing, managing and consuming the services of the future internet and managing technical and business issues.
- • Internet of Things: the nexus where Internet services interact and take advantage of the ubiquity of heterogeneous and resource-constrained nodes in the Internet of Things..
- Interface to Networks and Devices (I2ND): This layer provides the open interfaces to networks and devices.
- Security: This layer ensures that services are reliable and meet security and privacy requirements.

III. Wireless Sensor Networks (WSN)

Wireless Sensor Networks (WSN), together with RFID/NFC, are the primary building blocks of the Internet of Things. A WSN typically consists of a number of small sensor nodes that can sense and react to their surroundings, process information, and wirelessly communicate with one another; and one or more information receivers that receive and process the information reported by the sensor nodes, thereby controlling the WSN's operation. A WSN can be placed

in the environment to interact with the physical world, i.e. to supply us with real-time states of the physical world via rich sensory input as well as the ability to control the physical world via actuators. Sensor networks are critical components of the Internet of Things. A sensor network allows for quick access to information at any time and from any location. This function is carried out through gathering, processing, analyzing, and disseminating data [23]. WSN and RFID integration strengthens IoT and enables the development of IoT applications [24].

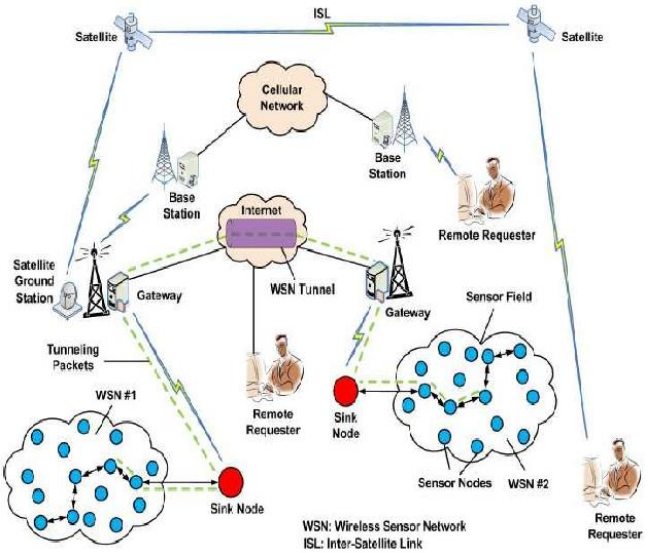


Fig 4: Structure of a WSN network**.**

A. Applications Of WSN

In the recent past, WSNs have found their way into a wide variety of applications and systems, with a wide range of requirements and special features. As a consequence, it is increasingly difficult to discuss the typical requirements regarding WSN hardware and software problems. This is especially problematic in a multidisciplinary research area, such as WSNs, where close collaboration between users, application domain experts, hardware designers, and software developers They need each other to implement effective systems in this area.

• Industrial: Monitoring and control of industrial equipment. Control of manufacturing processes and industrial automation. Manufacturing surveillance.

• Military: Knowledge of the situation in the combat field. Intruder detection in the facilities, movement detection units in enemy terrain, detection against chemical and biological threats and logistics control in urban warfare. Surveillance of the battlefield. Command, control, communications, computing, intelligence, surveillance, reconnaissance, and location systems. Military or civil assistance.

• Location of people and places.

• Low-rate wireless networks for precise location: Tracking assets, people, or anything that can move in different environments, including industrial, hospital, residential, and office environments.

• Monitoring: Supervise and control the physical world: The deployment of distributed sensor networks for a wide range of biological and environmental control applications, for marine, terrestrial and atmospheric environments; observation of biological, environmental and artificial systems; environmental monitoring of water and soil; discreet labeling of small animals, objects in a factory or hospital setting. Habitat monitoring (to determine the population and behavior of animals and plants). Maintenance of certain physical conditions (temperature, light, etc.).

• Civil protection: Detection and determination of the location of disaster sites. Detection of fires, earthquakes or floods.

• Automotive: Tire pressure control. Active mobility monitoring. Vehicle tracking. Traffic control.

• Airports: Smart labels. Wireless luggage tags.

• Agriculture: Sensors for soil moisture levels, pesticides, herbicides, pH. • Emergency situations: Control of levels of dangerous chemicals. Fire detectors. Monitoring of disaster areas.

• Machinery: Monitoring and control of machinery. Study of the movements of objects or structures. • Earthquakes or disasters: Warning systems.

• Commercial: Inventory management and product quality control.

• Health: Monitoring of patient locations and health conditions. Sensors for blood flow, respiratory rate, electrocardiogram, blood pressure, etc. Observation of patients. Helps patients with disabilities. Visual improvement of a patient with optic deficiency.

B. Routing in WSN

In a network, data routing determines the path or paths that data units (frames, packets, messages, and so on) will take from the source device to the destination. Most routing algorithms in interconnected networks allow you to choose paths that reduce data transmission latency. The routing devices exchange data in order to obtain a representation of the delays across the network set. Minimum delay routing helps to balance network load by lowering local congestion and connection delays. The routing algorithm is an essential component of any network. Wireless network routing algorithms must take into account the characteristics and constraints of this form of network [1]. Due to the characteristics that separate WSNs from other wireless networks such as ad

hoc mobile networks or cellular networks, routing in WSNs is a tough problem to solve. The difficulty of routing in WSN is primarily caused by the network's relatively large number of sensor nodes (hundreds to thousands), as it is not feasible to build a global routing scheme for the deployment of a large number of sensor nodes. It is also not recommended or practical to store the amount of data pertaining to the network's node identifiers in each node. Consequently, WSNs cannot use wired network protocols. A routing algorithm that satisfies the defined conditions, in accordance with the application's requirements, must be set in order for the WSN nodes to employ a cooperative methodology to establish communication. Since the sensors are typically powered by batteries and are frequently installed in locations where it is challenging to alter or replace them, any routing strategy proposed to function in WSNs must be efficient in terms of energy usage. The routes created by the routing algorithm should be chosen in accordance with a predetermined goal, such as maximizing network lifetime, guaranteeing that all data collected from the environment reaches the base station, or reducing network overhead. Network traffic, the amount of time it takes for data to move from the generating node to the base station, and so on. A good path is one with the lowest cost in respect to the aim indicated in any of the preceding questions. Sometimes the quickest path is not the most cost-effective. [18].

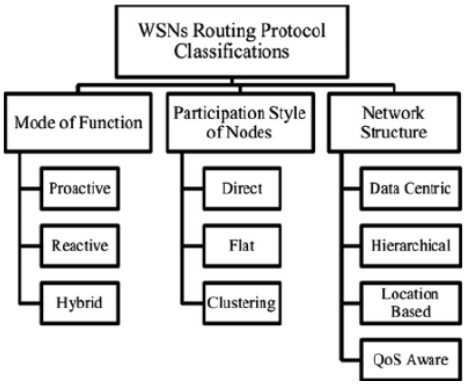


Fig 5 WSN routing protocols classification

IV. Machine Learning

Machine learning is a field of computational intelligence research that studies the development of methods capable of extracting concepts (knowledge) from data samples, which aims to build computational models that can adapt and learn from experience.

In general, the various ML algorithms are used in order to generate classifications/pattern recognition, modeling, simulations and numerical predictions for a set of samples. Therefore, ML techniques are used in induction (from a training set) of a classifier, which should be able (ideally) to predict the class of any instances of the domain in which it was trained.

In practical terms, machine learning algorithms aim to discover the relationship between the variables of a system (input/output) from sampled data. output) are fully understood. This is definitely not the case for many of the real problems we face in our daily lives.

Three paradigms can be used to generate a predictor through ML techniques: supervised, unsupervised and by reinforcement The choice of a learning paradigm determines the way in which the ML algorithm relates to its environment, that is, the way in which learning will occur through a set of data.

In supervised learning, there is an external intervention, which presents a set of examples in the form of input and output, in this method the algorithm is provided with a set of examples with known class labels, these examples being described as vectors of characteristic values or attributes and the associated class label. in supervised learning, we have a well-defined input and output and it is already known that there is a relationship between them. It should also be noted that supervised learning is divided into 2 methods: classification and regression Classification consists of problems in which the input data are already known and the algorithm has the task of assigning these data to certain classes

The regression method consists of a problem in which the input and output data are related through a continuous function, and the output of this model is normally a number

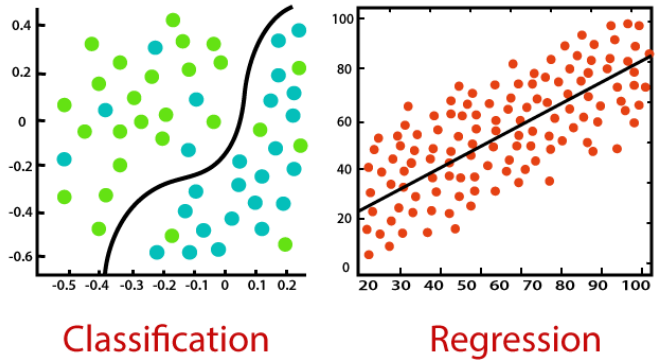


Fig 6 Classification and regression in supervised learning.

In unsupervised learning, there is no presence of this external intervention, forcing the algorithm to learn and represent the submitted inputs according to a quality measure. exposes that in this method only the input data is known and the main objective of this method is to find patterns or regularities in the input set. Unlike supervised learning, this method generates a cluster or clustering This process consists of grouping input data that have information in common or that are similar to each other,

forming different groups or different clusters. Clustering performs this grouping and analyzes the degree of similarities and differences between the various clusters generated in the process.

X. PROPOSED SYSTEM

A new research field emerged in climate science in the early 2000s that wanted to explore the increasing prevalence of extreme weather events like floods, storms, cyclones, etc. The field is known as "extreme event attribution" and has gained momentum in recent years in media in addition to the scientific world. There is mounting evidence that human activity is to blame for the increased risk of these extreme weather-type events. Researchers have also given importance to analyzing the economic costs linked to the human contribution to weather events. A study in 2020 approximated that nearly $67bn of damages caused by Hurricane Harvey in 2017 could attribute to human influences on climate. There are numerous methods to carry out attribution analysis. One way is to record instances of an extreme weather event and see their frequencies change with changes in environmental factors. We aim to build a model that accurately predicts the estimated damage to property while considering various event-related factors, in addition to external factors that might be influencing the extent of the damage.

A. Dataset

For this project, we have used publicly available data from the National Oceanic and Atmospheric Administration (NOAA) that contains event details on disaster incidents occurring in the US ranging from 1950 to August 2021. Some of the variables that we use from this dataset are as follows:

- begin and end date-time of event
- state where the event occurred
- the type of event (Hail, Storm, Drought, etc.)
- number of injuries and deaths
- starting and ending latitudes and longitudes of the event

We have also pulled in environmental indicators from yearly data collected by the United States Environmental Protection Agency (EPA). We have joined this data as additional information against the event year. The datasets that we have considered from the EPA source are as follows:

- emissions of greenhouse gases from 1990 to 2019
- events of heavy precipitation by land area percentage
- yearly earth surface temperature
- CSIRO and NOAA data for yearly sea-level changes
- variations in average seasonal temperature for fall, winter, summer, and spring
- artic ice coverage in March (yearly high) and September (yearly low)
- Glacier mass balance and number of observed glaciers

1. Linear Regression

It is a method to model a relationship between one or more independent variables and a response variable by fitting a linear equation on the observed data. Regression tells us the value of the response variable for an arbitrary explanatory variable value. The regression equation is:

$\hat{y} = b0 + b1\ x1 + b2\ x2 + \ldots$ where

b_0: intercept

b_i: slope/rate of change

2. Bootstrap aggregation

Bootstrapping is a sampling technique to create subsets of observations from the original data and is also known as bagging. In this technique, a generalized result combines the results of various predictive models. The subset size for bagging may be smaller than the original dataset.

3. Random Forest Regression

It is a supervised machine learning approach that solves regression and classification problems using bagging. The algorithm works by simultaneously training numerous decision tree estimators and outputing the mean or mode of all individual predictions. It prevents individual trees from overfitting the data and becoming locked in locally optimal solutions.

4. Extreme Gradient Boosting Regression

Gradient boosting is a class of ensemble machine learning algorithms constructed from decision tree models. It fits the model using any arbitrary differentiable loss function and gradient descent optimization algorithm. This technique is known as gradient boosting as we minimize the loss gradient while training the model.

Extreme Gradient Boosting, or XGBoost for short, is an efficient open-source implementation of the gradient boosting algorithm. XGBoost is a powerful approach for building supervised regression models.

B. Experimental Setup

1. Extraction

a. NOAA Data:

The NOAA data files are extracted from this link and have the following naming structure: StormEvents_details-ftp_v1.0_d1950_c20210803.csv.gz. The files are then concatenated into one to form our source data frame. We save this as a pickle file. This is done through get_NOAA_data() function.

b. EPA Data:

We use Pandas to read the various EPA data CSV files and collate them into one. We interpolate the missing data ranging back to the year 1950 by using the impute_EPA_data () function. We use the interp1d method from SciPy to get the extrapolated variable values.

Finally, we join the entire data into one data frame and remove the outliers from all the numerical variables.

1. Encoding categorical columns:

For categorical columns, based on type of data available, we did label encoding and one-hot encoding.

2. Imputation of logically important columns:

For column "DAMAGE_CROPS", we believed instead of simply removing all NAN's it is better to impute them with the average value of DAMAGE_CROPS per EVENT.

3. Split the data into training and validation sets:

Using from sklearn.
model_selection import train_test_split, able to create the training data and validation data sets.

4. **Standardize and normalize the data**:

Using,from sklearn.
preprocessing import StandardScaler, able to standardize the training data before running the regression models. In addition to this, using mean and standard deviation I normalized the training data.

C. Training Models:

1. Linear Regression:

Linear Regression constructs a linear model with coefficients w = (w1,..., wp) to minimize the residual sum of squares between the observed and anticipated targets in the dataset.

2. Random Forest:

A supervised learning algorithm that is based on the ensemble learning method and many Decision Trees. Random Forest uses a Bagging technique, so all calculations are run in parallel and there is no interaction between the Decision Trees when building them.

3. XGBoost Regressor:

Gradient boosting is a type of ensemble machine learning technique built from decision tree models. The loss function and gradient descent algorithm are used to fit models. This gives rise to the term "gradient boosting," because the loss gradient is minimized when the model is fitted. Extreme Gradient Boosting, or XGBoost, is a powerful approach for generating supervised regression models that is an efficient implementation of the gradient boosting algorithm.

4. Ensemble Model:
For ensemble learning, we've used the sklearn function "VotingRegressor". Simply put, this regressor uses individual model predictions and then averages them out to form a final prediction.

D. Additional options tried to increase model efficiency

Outlier Removal:

Using the Inter-Quartile Range method, I was able to identify the outliers and remove them. This method helped improve the R-square of the model by 5%.

Principal Component Analysis:

Principal Component Analysis, or PCA, is a very popular dimensionality reduction technique. PCA is trying to rearrange the features by their linear combinations. One characteristic of PCA is that the first principal component holds the most information about the dataset. The second principal component is more informative than the third, and so on

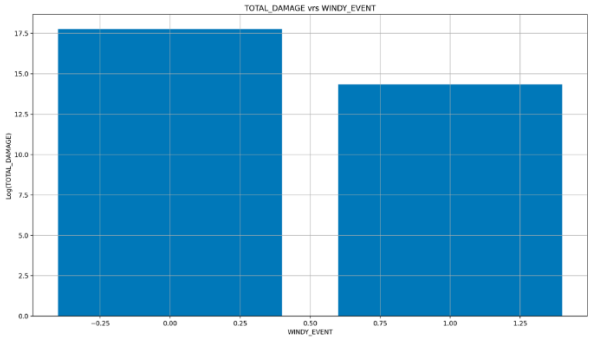
K-Means clustering for feature engineering

After multiple models tuning and feature creation iterations, the team could observe the increase in the model performance plateaued. The team decided to take help of the unsupervised K-Mean clustering model to create new feature hoping to increase the model performance.

XI. Results

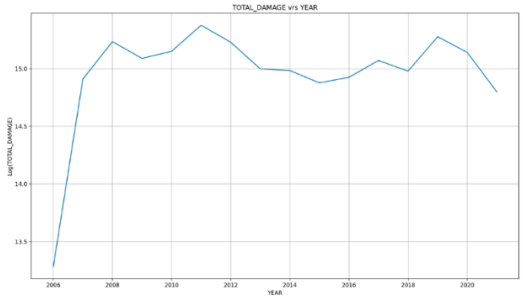
1. **Exploratory Data Analysis**

After obtaining the cleaned dataset, our objective was to get better insights about our data so that we can fix any data inconsistencies and get a clearer picture of event attributes that are explaining the variance in our target variable TOTAL_DAMAGE. We start by plotting the distributions of our target variable against various features to gauge their overall importance in our final model. For our plots, we take the logarithm of the total damage sum across different groups to plot our graphs. Here to plot our features with respect to target variable we have performed log transformation over our target variable.



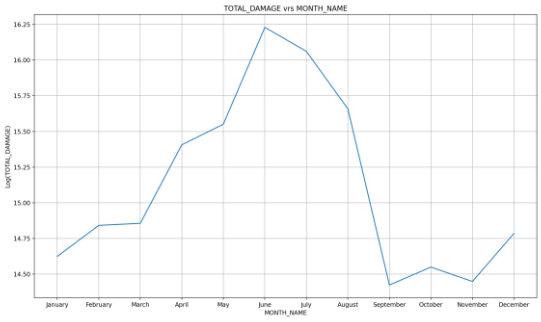
This graph depicts the total damage occurred due to a windy event when compared to a hail event

Plot 8: TOTAL_DAMAGE V/S YEAR



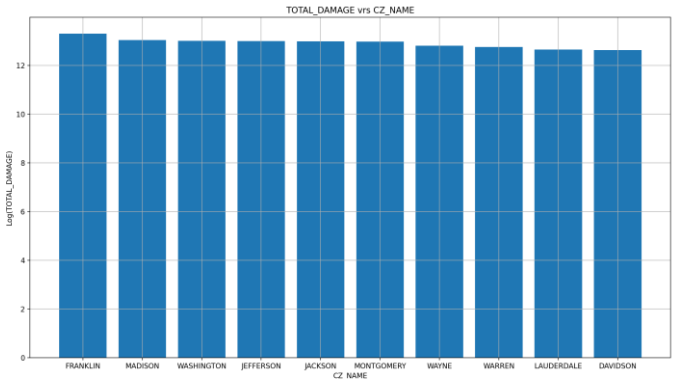
This graph shows the trend of total damage with respect to year. Here the graph is from 2006 (as before that NOAA didn't capture all the event types) which shows the increase in total damage with 2011 showing the highest total damage.

Plot 9: TOTAL_DAMAGE V/S MONTH_NAME



The months from May to August show the highest total damage as these months are the months for tornado season.

Plot 10: TOTAL_DAMAGE V/S CZ_NAME



The graph shows the top 10 counties with high total damage, Franklin being the highest.

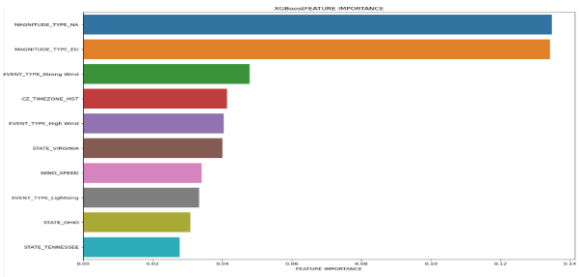
2. Modeling

We compare the performance of the different models by making use of the following metrics:

1. Mean squared error
2. Train R-squared value
3. Test R-squared value

2.1 XGBoost Regressor:

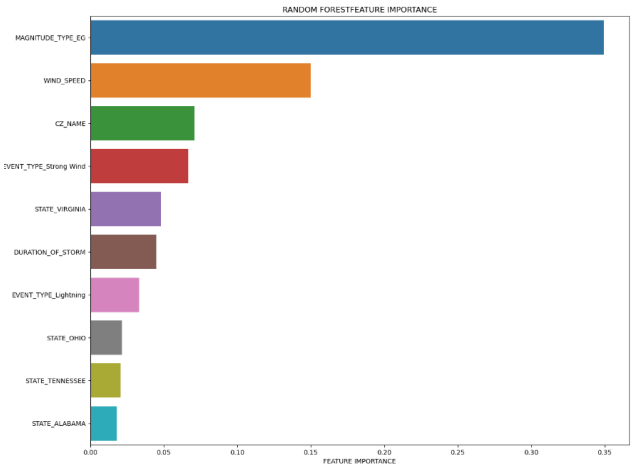
For XGBoost Regressor, our feature importance after training the model is given below. The RMSE values and train and test R-squared values are also included in the output.



We see that some of our important features from the model come out to be Magnitude_Type, Event_Type, State, CZ_Timezone, and Wind_Speed. Additionally, we get an R-squared value of around 44% which is not that great.

2.2 Random Forest:

For Random Forest Regressor, our feature importance after training the model is given below. The RMSE values and train and test R-squared values are also included in the output.



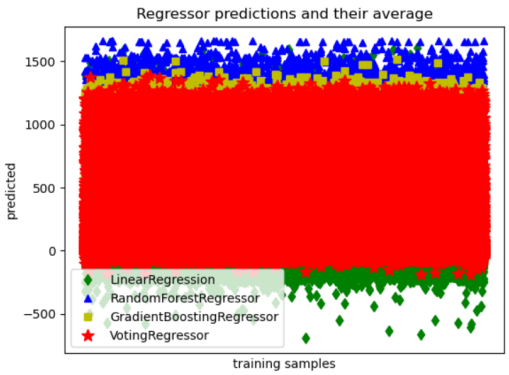
Similarly, we see that the important features are Magnitude_Type, Wind_Speed, State, Event_Type, Duration_of_Storm, and CZ_Timezone. We get better performance with R-squared value of about 50%.

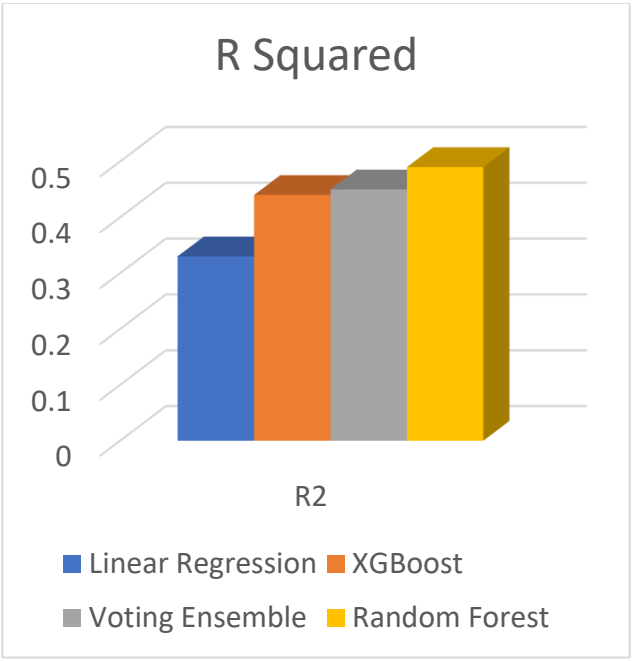
2.3 Rest of Models:

We also ran Linear Regression and Ensemble Model to predict the TOTAL_DAMAGE but both these models had a lower R-squared value than Random Forest.

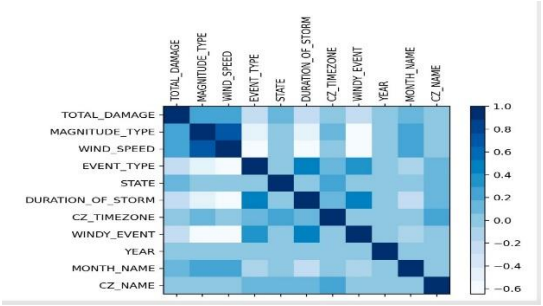
A detailed comparison of the models can be found below:

Model Name	Mean Square Error	Training R-squared Error	R-square error
Linear Regression	174992.1851	**33.70%**	**33.54%**
Random Forest	132421.8389	**52.93%**	**49.71%**
XGBoost	146355.2002	**45.22%**	**44.42%**
Ensemble	143642.2061	**44.39%**	**45.45%**



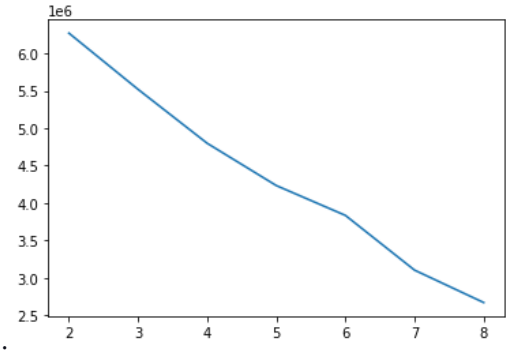


A correlation plot of our most important features against the TOTAL_DAMAGE variable returns the following result.



2.4. K-Means clustering for feature engineering

The team observed that after plotting the Sum of squared distances from the cluster mean for multiple number of clusters, the model could not provide a suitable number of clusters to use in the final model. Also the silhouette_score method resulted in inconclusive results due to the long run time of the model



The kmeans.inertia_ for the number of clusters

V. Conclusion

Both in terms of the number of lives lost and the amount of property that is damaged, natural catastrophes exact a heavy toll. Natural disasters exact a hefty toll. The proliferation of apps that make use of machine learning and deep learning has made it possible for scientists to keep up with the ever-increasing complexity of natural disasters, we are interested in learning about the ways in which these approaches have been implemented to improve the efficiency of various disaster management tasks. In order to accomplish this goal, we focused our research solely on recent publications and organized the results into categories according to the significance of the information they provided to various points in the sequence of events that led to the tragedy. In the context of these broader categories, machine learning and deep learning methods have been applied to a variety of natural disasters, including but not limited to floods, lava flows, earthquakes, typhoons, hurricanes, and landslides.

References

[1] Centre for Research on the Epidemiology of Disasters (CRED); United Nations Office for Disaster Risk Reduction (UNDRR). Global trends and Perspectives Executive Summary. 2021. Available online: https://www.undrr.org/publication/2020-nonCOVID-year-disasters (accessed on 4 October 2021).

[2] Altay, N.; Green, W.G. OR/MS research in disaster operations management. *Eur. J. Oper. Res.* **2006**, *175*, 475–493. [CrossRef]

[3] Sun, W.; Bocchini, P.; Davison, B.D. Applications of Artificial Intelligence for Disaster Management. *Nat. Haz.* **2020**, *103*, 2631–2689. [CrossRef]

[4] Drakaki, M.; Tzionas, P. Investigating the impact of site management on distress in refugee sites using Fuzzy Cognitive Maps. *Int. J. Disaster Risk Reduct.* **2021**, *60*, 102282. [CrossRef]

[5] Drakaki, M.; Gören, H.G.; Tzionas, P. An intelligent multi-agent based decision support system for refugee settlement siting. *Int. J. Disaster. Risk Reduct.* **2018**, *31*, 576–588. [CrossRef]

[6] United Nations Office for Disaster Risk Reduction (UNDRR). *UNISDR Terminology on Disaster Risk Reduction*; UNISDR: Geneva, Switzerland, 2009. Available online: https://www.unisdr.org/files/7817_UNISDRTerminologyEnglish.pdf (accessed on 4 October 2021).

[7] EM-DAT—The International Disasters Database. Guidelines. EM-DAT—Data Entry—Field Description/Definition. Available online: https://www.emdat.be/guidelines (accessed on 4 October 2021).

[8] Van Wassenhove, L.N. Blackett memorial lecture humanitarian aid logistics: Supply chain management in high gear. *J. Oper. Res. Soc.* **2006**, *57*, 475–489. [CrossRef]

[9] Arinta, R.R.; Andi, E.W.R. Natural disaster application on big data and machine learning: A review. In Proceedings of the 2019

4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), Yogyakarta, Indonesia, 20–21 November 2019; Volume 6, pp. 249–254. [CrossRef]

[10] Yu, M.; Yang, C.; Li, Y. Big data in natural disaster management: A review. *Geosciences* **2018**, *8*, 165. [CrossRef]

[11] Schmidhuber, J. Deep Learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef]

[12] Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]

[13] Presa-Reyes, M.; Chen, S.C. Assessing Building Damage by Learning the Deep Feature Correspondence of before and after Aerial Images. In Proceedings of the 2020 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), Shenzhen, China, 6–8 August 2020. [CrossRef]

[14] Akshya, J.; Priyadarsini, P.L.K. A hybrid machine learning approach for classifying aerial images of flood-hit areas. In Proceedings of the 2019 International Conference on Computational Intelligence in Data Science (ICCIDS), Chennai, India, 21–23 February 2019. [CrossRef]

[15] Fan, C.; Wu, F.; Mostafavi, A. A Hybrid Machine Learning Pipeline for Automated Mapping of Events and Locations from Social Media in Disasters. *IEEE Access* **2020**, *8*, 10478–10490. [CrossRef]

[16] Ben-Hur, A.; Horn, D.; Siegelmann, H.T.; Vapnik, V. A support vector clustering method. In Proceedings of the 15th International Conference on Pattern Recognition. ICPR-2000, Barcelona, Spain, 3–7 September 2000. [CrossRef]

[17] Hofmann, T.; Schölkopf, B.; Smola, A.J. Kernel methods in machine learning. *Ann. Statist.* **2008**, *36*, 1171–1220. [CrossRef]

[18] O'Connor, J.; Eberle, C.; Cotti, D.; Hagenlocher, M.; Hassel, J.; Janzen, S.; Narvaez, L.; Newsom, A.; Ortiz-Vargas, A.; Schuetze, S.; et al. Interconnected Disaster Risks. *UNU-EHS* **2021**, 64. Available online: https://reliefweb.int/report/world/interconnecteddisaster-risks-20202021 (accessed on 4 October 2021).

[19] Dwarakanath, L.; Kamsin, A.; Rasheed, R.A.; Anandhan, A.; Shuib, L. Automated Machine Learning Approaches for Emergency Response and Coordination via Social Media in the Aftermath of a Disaster: A Review. *IEEE Access* **2021**, *9*, 68917–68931. [CrossRef]

[20] Yuan, C.; Moayedi, H. Evaluation and comparison of the advanced metaheuristic and conventional machine learning methods for the prediction of landslide occurrence. *Eng. Comput.* **2020**, *36*, 1801–1811. [CrossRef]

[21] Sankaranarayanan, S.; Prabhakar, M.; Satish, S.; Jain, P.; Ramprasad, A.; Krishnan, A. Flood prediction based on weather parameters using deep learning. *J. Water Clim. Chang.* **2019**, *11*, 1766–1783. [CrossRef]

[22] Huang, Y.; Jin, L.; Zhao, H.S.; Huang, X.Y. Fuzzy neural network and LLE algorithm for forecasting precipitation in tropical cyclones: Comparisons with interpolation method by ECMWF and stepwise regression method. *Nat. Hazards* **2018**, *91*, 201–220. [CrossRef]

[23] Asim, K.M.; Martínez-Álvarez, F.; Basit, A.; Iqbal, T. Earthquake magnitude prediction in Hindukush region using machine learning techniques. *Nat. Hazards* **2017**, *85*, 471–486. [CrossRef]

[24] Amin, M.S.; Ahn, H. Earthquake disaster avoidance learning system using deep learning. *Cogn. Syst. Res.* **2021**, *66*, 221–235. [CrossRef]

[25] Prasad, P.; Loveson, V.J.; Das, B.; Kotha, M. Novel ensemble machine learning models in flood susceptibility mapping. *Geocarto Int.* **2021**, *26*, 1892209. [CrossRef]

[26] Nsengiyumva, J.B.; Valentino, R. Predicting landslide susceptibility and risks using GIS-based machine learning simulations, case of upper Nyabarongo catchment. *Geomat. Nat. Hazards Risk* **2020**, *11*, 1250–1277. [CrossRef]

[27] Shirzadi, A.; Bui, D.T.; Pham, B.T.; Solaimani, K.; Chapi, K.; Kavian, A.; Shahabi, H.; Revhaug, I. Shallow landslide susceptibility assessment using a novel hybrid intelligence approach. *Environ. Earth Sci.* **2017**, *76*, 60. [CrossRef]

[28] Sriram, L.M.K.; Ulak, M.B.; Ozguven, E.E.; Arghandeh, R. Multi-Network Vulnerability Causal Model for Infrastructure CoResilience. *IEEE Access* **2019**, *7*, 35344–35358. [CrossRef]

[29] Wahab, A.M.; Ludin, A.N.M. Flood vulnerability assessment using artificial neural networks in Muar Region, Johor Malaysia. *IOP Conf. Ser. Earth Environ. Sci.* **2018**, *169*, 012056. [CrossRef]

[30] Mutlu, B.; Nefeslioglu, H.A.; Sezer, E.A.; Ali, A.M.; Gokceoglu, C. An experimental research on the use of recurrent neural networks in landslide susceptibility mapping. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 578. [CrossRef]

[31] Shanthi, D. N. ., & J, S. . (2021). Machine Learning Architecture in Soft Sensor for Manufacturing Control and Monitoring System Based on Data Classification. Research Journal of Computer Systems and Engineering, 2(2), 01:05. Retrieved from https://technicaljournals.org/RJCSE/index.php/journal/article/view/24

[32] Anupong, W., Yi-Chia, L., Jagdish, M., Kumar, R., Selvam, P. D., Saravanakumar, R., & Dhabliya, D. (2022). Hybrid distributed energy sources providing climate security to the agriculture environment and enhancing the yield. Sustainable Energy Technologies and Assessments, 52 doi:10.1016/j.seta.2022.102142