3. FUNDAMENTALS OF THE IoT AND DEEP LEARNING

Internet of Things(IoT) changed the monitoring paradigm of air pollution from being dependent on government agencies that handled fixed and costly air quality monitoring stations with active human interaction to the low cost and auto functioning moveable monitoring systems. With the development of technology and research in the economical sensing technology, number of researchers are participating to the research related to air quality monitoring. The overall internet of things architecture and the ecosystem is discussed in detail here.

3.1 INTERNET OF THINGS ECOSYSTEM

Internet of Things is the technology that attaches various things (devices) and enables them for interaction. IoT technologies have the potential to mitigate the adverse effect of air pollution by availing the updates regarding the sudden fluctuations and status of air quality parameters to the citizens, environmental experts, and the government in near real-time. The IoT field is facing challenges as there is no standardization of protocols and architecture. The IoT architecture is described with different perspectives (three, four, and five-layer) in various literature, can be given below.

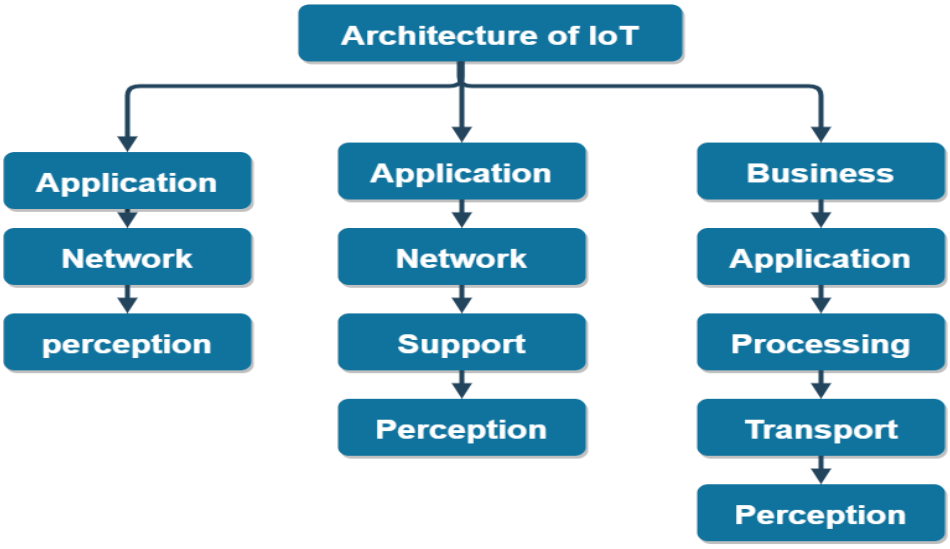


Figure 3.1. Layered IoT architecture with different perspectives

In general, the Internet of Things layered design consists of a physical layer, network layer, information processing layer, and application layer. A physical layer consists of a variety of sensors and actuators. The physical layer or perception layer is responsible for data collection, specific to the domain of study, and contains various sensors, sensor modules, and development board, enabling the physical communications. The physical layer provides the collected sensing data to the network layer; the network layer gathers sensor data and communicates it to the information processing layer via numerous mediums such as the Internet, Wi-Fi, Bluetooth, RFID, NFC, and many more. The information processing layer is responsible for analyzing sensor data and passed it on to the application layer for data representation. The application layer will represent analyzed data in the form of a web and mobile interface.

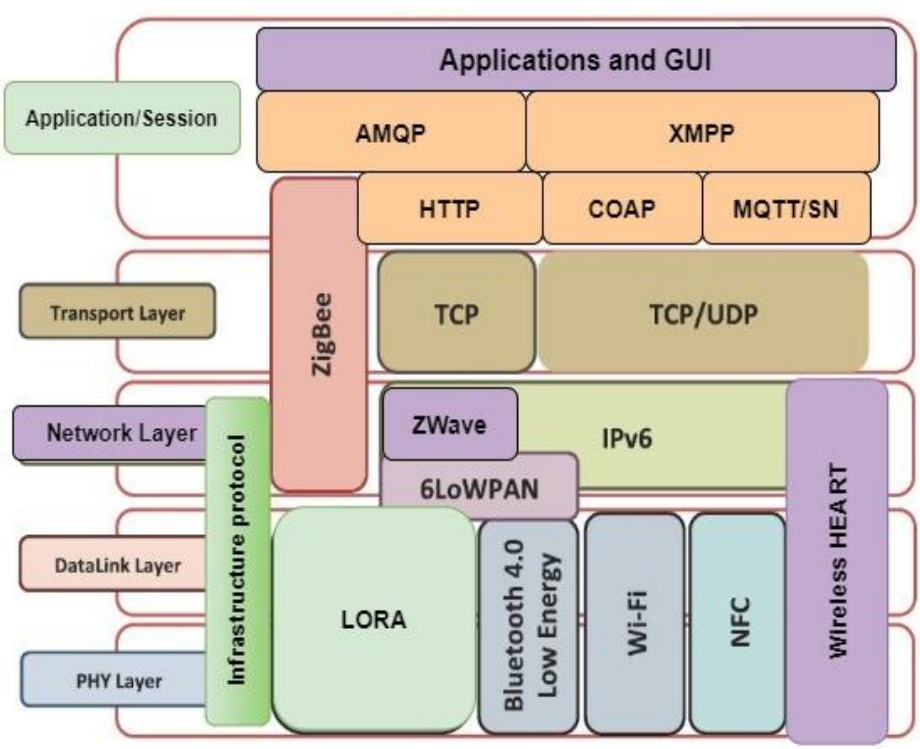


Figure 3.2. Layer wise protocol categorization when compared to ISO model

There are many groups and forums such as the World Wide Web Consortium (W3C), Internet Engineering Task Force (IETF), Organization for the Advancement of Structured Information Standards(OASIS), Institute of Electrical and Electronics Engineers(IEEE), and European Telecommunication standards institute (EISI) working for standardization of various protocols of IoT. The overall protocols involved in the IoT ecosystem are depicted in Figure 3.2. These protocols span multiple layers of TCP/IP architecture compared to the standard ISO

architecture for reference purposes. Still, there is no clear division of protocols/according to the layers involved in IoT architecture. The network layer could be combined with the data link and transport layer sometimes. Protocols related to physical, link, and network layers can be considered or categorized as infrastructure-level protocols, whereas the protocols related to the standard session or application layer could be merged in another classification or category.

3.1.1 Infrastructure Protocols

1. IEEE 802.11

The IEEE 802.11, known as Wi-Fi also, is one of the broadly accepted standard based on wireless communication and most digital devices accept the standard. All variations of IEEE 802.11 using various width for the channels. The IEEE 802.11 ah is established by IEEE 802.11 working group seeing a reduction in frame overhead and power optimization requirement appropriate for IoT related applications [105, 106].

2. Near Field Communication(NFC)

NFC [107] is one of the short range based communication based on wireless technology. Devices of NFC interconnect with each other by generating an electromagnetic type field. Near Field Communication can read and write data kept in five types of different tags. Earlier, the standard was not incorporating the security features. In the year 2015 reform to signature record definition is presented by specifying RTC 2.0 to ensure the authentication.

3. 6LoWPAN

The 6LoWPAM- IPv6 over low power wireless personal area network practices a compact header with IPv6 addressing. 6LoWPAN compresses the lengthy header of IPv6 in smaller frames with a maximum possible size of 128 bytes with the use of various options of the compression scheme. 6LoWPAN enables low bandwidth, fragmentation alternatives, mobility, compression, and feature of multi-hop routing by adding layer over datalink layer [108, 110].

4. Z-Wave

Z-Wave is also a wireless protocol with low-power consumption. The protocol was designed by the Sigma Designs, Inc., mainly for purpose of home automation. Z-Wave working is constructed on a master and slave kind of architecture. It uses ACK messages for providing reliability [109].

5. ZigBee

ZigBee is one more technology based on wireless communication, developed by ZigBee Alliance. ZigBee is extensively applied for sensor network building and WSN applications. The protocol stack of ZigBee is made-up over the IEEE 802.15.4. ZigBee offers two profiles, ZigBee and ZigBee Pro. ZigBee delivers the security, inherited from IEEE 802.15.4. IEEE 802.15.4 take cares of Physical and MAC layer functioning [111].

6. Long Range(LoRa)

LoRa is one of the low power and low-cost standard, providing a scalable alternative. LoRa practices a license-free frequency band of sub-GHz and provides the long-distance communication around ten km [112]. LoRa provides the security features, device authentication at the network level, and integrity related check and encryption at the application level.

Table 3.1. Comparison of various infrastructure layer protocols

Proto. Name	Range	Frequency	Data Rate	Network Topology	Security
IEEE 802.11	Up to 100m	2.4 or 5 GHz	11Mbps(802.11b), 54Mbps(802.11a/g), 450 Mbps(802.11n)	Point to point, Point to Hub	Encryption WEP,WPA, WPA2
NFC	4-10 cm	13.56 MHz Radio frequency	424 Kbps	Point to Point	LES, AES
BLE	100 m	2.4 GHz	2 Mbps	Point to point, small n/w	128-bit AES
6LowPAN	10-100 m	902-929 MHz(US), 2.4 GHz(world)	250 Kbps(2.4 GHz), 40 Kbps(915MHz)	Mesh	128- bit AES
Z-Wave	100 m	865.2MHz(India), 868.42(EU), 908.4(US)	100 Kbps	Source routed Mesh	128- bit AES
ZigBee	300 m	2.4GHz(world) ,900 MHz(US) ,868 MHz(EU)	250 Kbps	Star, Mesh	128- bit AES
LoRa	10 km	868 MHz (EU), 923 MHz (Asia)	0.25 Kbps – 27	Star, Mesh	128- bit AES

3.1.2 Session/Application Protocols

1. Message Queue Telemetry Transport(MQTT)

MQTT is an open ISO standard lightweight application level protocol. The protocol operates over the underlying TCP/IP. The protocol works on Pub-Sub (publication-subscriber) kind of form of working. In 1999, Andy Stanford at IBM and Arlen Nipper at Cirrus Link authored the very first version. Later, IBM in 2013 established a third version of open standard and delivered it to OASIS [113]. The motive of the protocol is to interconnect low capacity embedded devices over a network or internet for the communication purpose at resource-constrained places or sites. The protocol will be discussed in detail under the system implementation section.

2. MQTT for Sensor Network(MQTT-SN)

MQTT-SN is one of the optimization and extension of the basic MQTT protocol. It is altered to suit low-power devices such as sensors and actuators, having limited capacity for processing and storage. MQTT utilizes topic names, whereas MQTT-SN practices topic ID to reduce transmission overhead [114]. MQTT-SN supports sleep mode at the client side. Stored messages at the broker can be retrieved later by clients on waking up. The client gets connected to the integrated gateway at the broker. If the gateway is not integrated, the client connects to the gateway in WSN and connects with the broker over MQTT. Gateway transforms MQTT-SN messages into simple MQTT packets. Two connection types of support are available between the gateway to broker, transparent and aggregating. In transparent mode, each client has its connection to the broker, whereas a single connection is used in aggregated mode.

3. Constrained Application Protocol(COAP)

Motes prepared with controllers have limited memory for processing, and IPv6 networks over 6LowPan have a low throughput rate. CoAP is a web communication related protocol specifically intended for devices with low power facility and is connected to lossy networks. CoAP was developed by Internet Engineering Task Force(IETF) and ARM, and it was standardized in the year 2014. CoAP works on request-response-based architecture. CoAP is using UDP as a transportation layer protocol for asynchronous message exchange mechanisms [115].

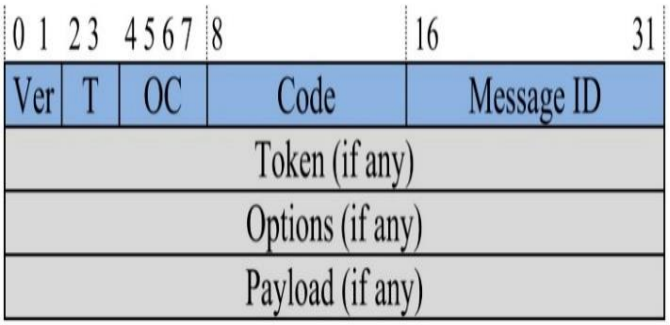


Figure 3.3. The message format for the CoAP protocol

CoAP is more appropriate for IoT-based applications due to UDP support at the transportation layer. CoAP requires only four-byte for the header at the application layer, and UDP also needs only an eight-byte packet header. As shown in Figure 3.3, two bits of the version field define the COAP version. In CoAP, various types of messages are processed such as confirmable, non-confirmable, reset and the acknowledgment. The field - type defines the type of the message. The MessageID is required and utilized to synchronize the messages and identify messages counter to the acknowledgment or reset. The client produces a token field to maintain the uniqueness of the source and destination pair. The option is the field used for switches/options implemented by the protocol. If the option field indicates a critical category, then the receiver entity must understand all the field values and act; otherwise, it is elective. The receiver can ignore the field values. CoAP is the protocol having support for multicasting because it is built on IPv6. CoAP works on Representational State Transfer(REST) architecture, shared with HyperText Transfer Protocol(HTTP) and implements get, put, post and delete methods inherited from HTTP [115, 116]. The protocol supports service as well as resource discovery. The resource discovery aims to deliver the list of resources and the respective URI with which resource is registered at the server. With Datagram Transport Layer Security(DTLS), CoAP avails security alternatives, and the implementation is known as secure CoAP (CoAPs). CoAP is still in the development stage, and very little open forum support is available for customization and implementation.

4. Advanced Message Queuing Protocol(AMQP)

AMQP is an OASIS standard established for business and financial industry purposes. AMQP also works on pub-sub(publish-subscribe) pattern-based architecture, similar to the MQTT protocol. The underlying transport protocol used is TCP. TCP is responsible for reliable communication. AMQP uses brokers as an intermediary entity between the publisher and subscriber. The broker entity is comprising of two modules, queues and exchange. Queues are

used as storage and forwarder. The messages published by publishers are received at the exchange, and then messages are transferred to the various queues based on pre-specified rules and conditions.

5. Extensible Messaging and Presence Protocol (XMPP)

Extensible messaging and presence protocol (XMPP) is a protocol intended for instant messaging and scalability. The protocol exploits real-time messaging for authentication, access control, and confidentiality. XMPP is using Extensible Mark-up Language (XML) data stream for asynchronous messaging between client and server. XMPP is a client-server communication protocol. However, some extensions of the protocol are available that make it execute the pub-sub model. XML stanza, which is structured data that is exchanged amongst nodes. The XML stanzas can be shifted between all the clients accessing the same server. The Clients are connected to the server using IP address and port number maintained by underlying TCP protocol. For delivery to the destination, the protocol utilizes its addressing and address resolution scheme. The stanzas are also exchanged between servers as per the pre-defined terms. The transfers of stanzas between servers create inter-domain communication between connected clients. The protocol is also enabled with built-in TLS support for encryption and integrity preservation. The TLS support makes the protocol secure compared to the other protocols. XMPP uses XML for messages, and due to that, the size of the messages is one of the challenges in applications with constrained bandwidth networks. This protocol is not feasible for low energy and lossy type of networks due to the persistent TCP connection requirement [117]. Table 3.2 represents the comparison of the session/application level protocols over different parameters.

Table 3.2. Comparison of session/application layer protocols

Protocol	**XMPP**	**AMQP**	**COAP**	**MQTT**
Standard	OPEN	OASIS	IETF	OASIS
Transportation protocol	TCP	TCP	UDP	TCP
RESTFUL	NO	NO	YES	NO
Identifier	URI	URI and Message topics(both support)	URI	Topics

Security	TLS	DTLS	DTLS	SSL
Header Size(Bytes)	-	8	4	2
Communication Pattern	Publish-Subscribe	Publish-Subscribe and point to point both	Request-Response	Publish-Subscribe

3.1.3 Low-Cost Sensors for Air Pollutant Gases

The air quality index(AQI) is created and utilized by the governments or agencies to specify the overall level of air pollution calculated from the measured levels of various individual air pollutants. Major air pollutants identified considering the pollutants included in the air quality index are ozone, various oxides of carbon, sulfur and nitrogen, and particulate matter [118, 119]. Commercial sensors for sensing gaseous pollutants can be categorized into four major categories as per their working principle and technologies. The categories are electrochemical, resistive, photoionization, and dispersive infrared radiation absorption (NDIR) [120].

Resistive sensors use resistance as for the measurement of the nearby gases. The operation of these sensors is built on numerous sensing and transduction factors or parameters. Recently, the resistive gas sensors having metal oxide semiconductors (MOS) and conducting polymers layer are widespread and discovered space looking at their application for calculating the gas pollutants. The MOS active layer alters the resistance along with the gas exposure. The variation in the measurements having relation to the concentration of the gas [121, 122]. MOS-based sensors are small, lightweight, economical, and also have a fast response time. These sensors need high temperatures for a quick reaction rate [120]. These sensors undergo cross-sensitivity issue and are having a non-linear response curve [123].

Electrochemical sensors (amperometric) work is based on the reaction that emerges between the target gas and the electrolyte. Such a reaction produces the current and the produce current reflects the quantity of concentration of gases [120]. The Electrochemical sensors are comprised of three electrodes infused with electrolyte material. The electrolysis is experienced by targeted gas at one of the three electrodes (working electrode). The working electrode is usually maintained on a gas-permeable membrane. Another electrode, the counter electrode, handles the balance of the current generated by the working electrode. The observed electric current matches the concentration of the gas. The electrochemical sensors are also facing the

issue of cross-sensitivity, but the problem can be controlled by selecting various available electrode materials. Electrochemical sensors are more practical to use in IoT-based solutions because of the lower power requirement. Usually, they deliver linear output in comparison to the non-linear response generated by MOS sensors. These sensors are having issue of measuring low concentration range detection [120,124]

A non-dispersive infrared sensor technology is an optical transducer-based technique with the use of infrared gas absorption. Vapours can absorb infrared radiation from Volatile Organic Compounds (VOC) and some flammable gases. The targeted gas absorbs the specific wavelength radiation. Due to the absorption, radiation intensity decreases, and the decrease in radiation is converted into an electrical signal equivalent to the gas concentration [125].

In the photoionization measurement approach, the target gas molecules are irradiated by the ultraviolet light of high energy. The process breaks the molecule and produces electrically charged ions. The produced charged ions are exposed to the electrical field to produce a current that is proportionate to the gaseous concentration [120,125]. Table 3.3 displays a summary of significant categories as per the operating principles, with strengths and concerns. Table 3.4 gives commercially available and frequently used gas sensors along with specifications from well-known manufacturers.

Table 3.3 Summary of significant categories (working principles) of low-cost sensors

Sensing Technology	**Curve**	**Gases**	**Power Consumption**	**Strengths**	**worries**
Electrochemical	Linear	Ozone, CO, NO_X	Very Low	Low cost, small size	Sensitivity to other gases, Challenging to measure low concentration range
Photoionization	Near Linear	Ozone	Medium	High accuracy, stability, Very quick response time	Not compact
NIDR	Non-Linear	VOC, CO_2	Low	Stability against humidity and temperature	Path length affects performance
MOS	Non-Linear	Ozone, CO, NO_X	Medium	Low cost, small size, stable	Sensitive to variation of relative humidity and temperature, Cross sensitivity issue,

					Continuous heating/ warm-up time requirement

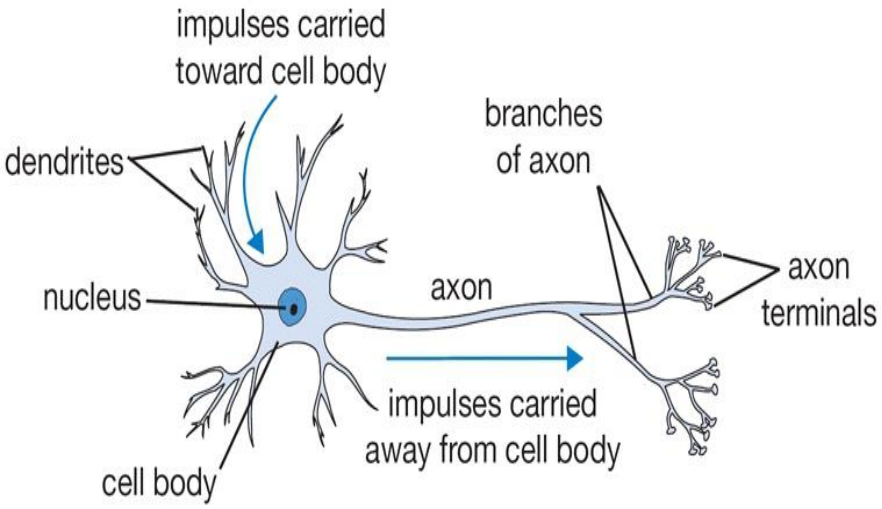
Table 3.4 Commercially available and frequently used gas sensors

Target Gas	Sensor	Manufacturer	Type	Detection Range	Specification
CO	FECS40	FIGARO	EC	0-1000 ppm	Resolution: 1ppm, Response time <30 sec Expected life time: 3years, Weight: 4.5g Operating T: -20~50 C , Operating RH : 15~ 90%
CO	TGS3870	FIGARO	MOS	50-1000 ppm	Heater vol.: 0.9V±3% (High temp.).,0.2V±3% （Low temp.） Heater resistance: 1.8kΩ~24kΩ
CO	ME2CO	WINSEN	EC	0-1000 ppm	Resolution: 0.5 ppm, Response time <50 sec Expected life time: 5years Operating T: -20~50 C, Operating RH : 15~ 90%
CO	ZE07	WINSEN	EC	0-500 ppm	Resolution: 0.1ppm, Response time <60 sec Expected life time: 2years, Weight: - Operating T: -10~55 C ,Operating RH : 15~ 90%
CO	MQ-7B	WINSEN	MOS	10-500 ppm	Heater voltage:5.0V±0.1V AC or DC (High temp.） 1.5V±0.1V AC or DC (Low temp.） Heater resistance: 29Ω±3Ω
CO	MICS 5525	E2V	MOS	0-1000 ppm	Heater voltage: - Heater resistance: 74±8Ω

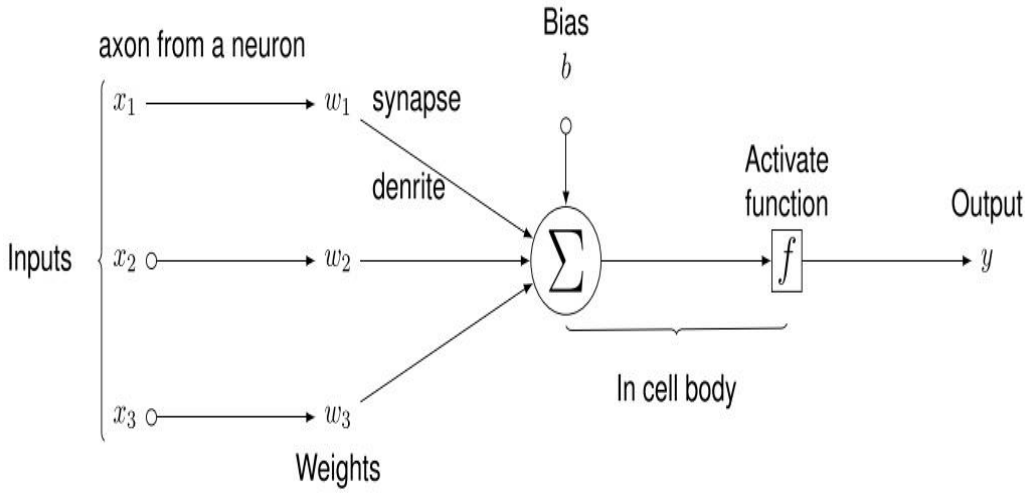
3.2 DEEP LEARNING

Deep learning is the mechanism typically termed an artificial neural network (ANN). ANN with one or more hidden layers by the hierarchical learning process empowers the computer or machine to excerpt high-level and complex abstractions as data or feature representations. Deep learning excludes the necessity of handmade feature generation, typically derived using expert knowledge of the domain. The hierarchy-based learning structure of the deep learning methods is inspired by artificial intelligence matching the deep and layered learning of the significant sensory parts of the neocortex available in the human brain, which can automatically fetch features and abstractions from the underlying data given.

The neuron (refer to Figure 3.4 (a)) is the fundamental calculation unit in our human brain. The neuron obtains the input signals from dendrites, and in turn, it generates output signals along the axon. The axon is further connected to the dendrites of similar other neurons via synapses. The output signals transmitted along the axons can communicate with the dendrites of those other neurons; the communication or interaction depends on the synaptic strength.



(a)



(b)

Figure 3.4 (a) biological neuron structure (b) mathematical model of the neuron

Figure 3.4 (b) shows the mathematical model representing a similar mechanism to the biological neuron network. The signal that traverses along the axon is represented as x, and the synaptic strength is denoted as w. The communication or interaction at the synapse is modelled

as the multiplication of x and w. The assumption is that all such signals passed by are added in the cell body. Also, one more assumption is that if the summation is more than a particular threshold value, the neuron can fire with the transfer of spike along the axon. The frequently utilized activation function is the sigmoid σ. The sigmoid activation function gets a real value as input (strength of the signal after the addition). The sigmoid activation function maps the input to the output in the range [0, 1]. In this mathematical neuron model, the core idea is to learn the weight w- synaptic strength and regulate the strength of the effect or influence like the directions: executory –with positive weight or inhibitory- with the negative weight of one neuron on one neuron another.

Neural networks (NNs) can be represented as a group or collection of interconnected neurons connected in an acyclic graph fashion (no directed cycles). The neural networks are structured as layers of neurons. The neurons from one layer are connected to the neurons from the adjacent or next layer, but the neurons from the same single layer are not sharing any connection. An example of such a network is shown in Figure 3.5. The input layer is the first one, which is designed and dimensioned as per the input. This layer is the interface between the input data and the neural network. The next layer is the hidden layer, which contains several connected neurons responsible for processing information. The output layer is the last one representing the result of information processing in the network.

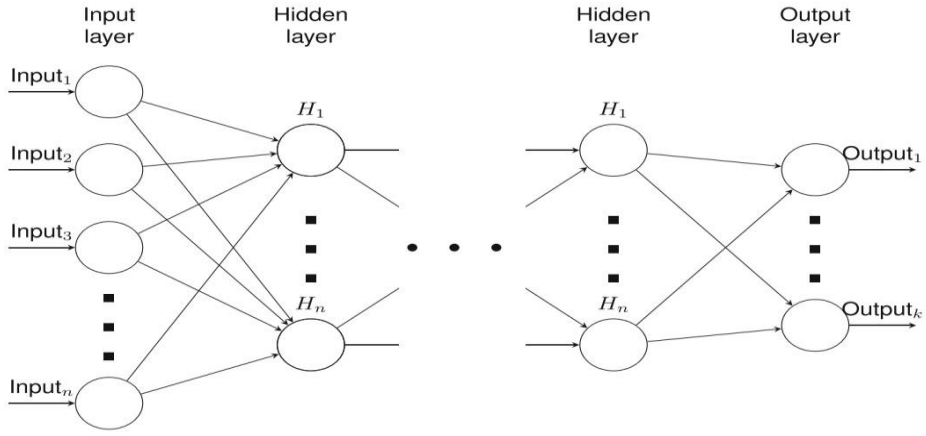


Figure 3.5 An example of a neuron network

The neural networks can be categorized into feedforward and feedback neural networks based on the interconnection between the hidden layer neurons. The feed-forward is the neural network in which the information traverses in a single input to output or forward direction. The outcome of one hidden layer neuron is conveyed as the neuron's input to the next coming

hidden layer. There are not any loops of connection or information present in such a network. In a feedback neural network, the information is passed in both directions. The neuron can be connected to any other neuron in the hidden layer without the need for hierarchy, and thus the underlying structure allows a loop in the network. Such a nature allows information to traverse continuously and alter the network parameters until the optimized or state of equilibrium is not reached.

From the time of inception, the deep neural network is effectively used in fields like image classification, natural language processing, gesture and object recognition, decision support and recommendation systems, and biomedical informatics. Various structures of deep learning networks comprise deep Convolutional Neural Networks, Deep Recurrent Neural Networks, Deep Multi-Layer Perceptions, and Deep Auto Encoder Networks.

3.2.1 Recurrent Neural Network

Among these all architectures, DRNN is principally appropriate for time series predicting and modelling. DRNN or simply RNN (recurrent neural network) is one of the types of the feedback network. Feed-forward neural networks do not have any memory to remember the state of the input they applied with so weak in predicting the next state. As the feed-forward neural network studies the current input only, the network is missing the concept of order in time. The feed-forward network merely cannot recollect anything regarding what occurred in the past apart from the training. While in the recurrent neural network, the information traverses through the loop. While RNN takes the decision, RNN considers the current input and considers the trace of past gathered information using recurrent connections as shown in the below figure 3.6.

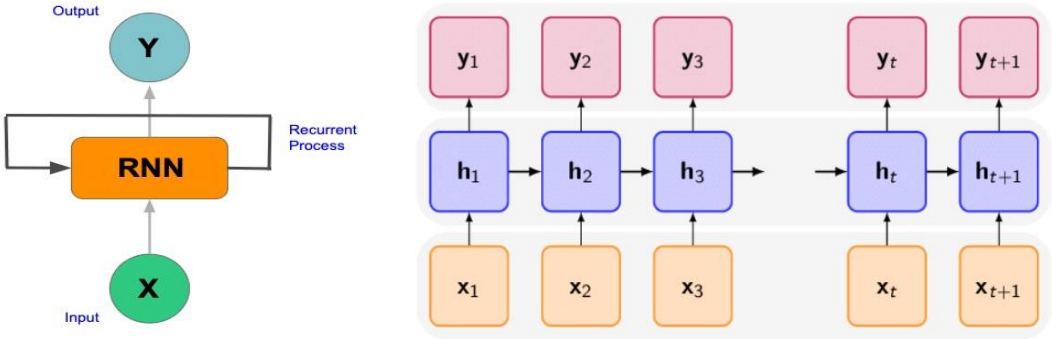


Figure 3.6 Recurrent neuron network unrolling

The input time series data given to the RNN is $X = (X_1, X_2, \ldots, X_t)$; here, X_t is the input at time step t. RNN will recursively process the input sequence data to learn the representation

of the inherent pattern repeatedly that appeared in the past. With this operation consideration, RNN preserves internal hidden state h_t while processing the input sequence and share it across all the time steps

$$h_t = f(h_{t-1}, x_t) \tag{3.1}$$

where h_t is the new state, h_{t-1} is the previous state, and x_t is the current input. Here, the new state is dependent on the current input and the previous input state because the input neuron has applied transformation on the previous one. Hear each successive input is termed as the time step. Considering the simplest form of RNN and hyperbolic tangent as an activation function, if the recurrent neuron weight is W_{hh} and the current input neuron is $W_{xh,}$ then the equation for state t can be given as equation 3.2.

$$h_t = tanh(w_{hh}h_{t-1} + w_{xh}x_t) \tag{3.2}$$

Various functions can be utilized as activation functions such as sigmoid, tanh, relu, etc.

Sigmoid Function: The value of the function inclines to zero when the input variable says z approaches to negative infinity value and approaches to 1 when the input variable z approaches to infinity. So the sigmoid function provides output in the range [0,1].

Tanh Function: The tanh function is the rescaled form of the sigmoid, and the output range is replaced with $[-1,1]$ in place of the range [0,1] in the sigmoid.

ReLU Function: The Rectified Linear Unit(ReLU) is a frequently used function in deep learning. The function gives 0 as output if it gets any negative input. For any positive value x, the function gives x as output. So, the behaviour can be given as f(x)=max (0, x).

Finally, the SGD- Stochastic Gradient Descent and BP- Back Propagation algorithms are utilized for training the neural network function and getting the optimal parameters. In a neural network, the forward propagation is done to get the output and test if the output is correct in terms of error. The backpropagation is going back through the network for deriving the partial derivatives of related error with the applied weights; the process enables the reduction of the values of weights. The partial derivatives concerning the inputs are known as the gradient. The gradient measures the amount of output of function changes if there is a bit change in inputs. The gradient is like a slope function. The slope is steeper as the gradient value is higher, and the model can learn quickly. It calculates the change in all the weights with respect to the change of error. These derivatives are then utilized by some gradient descent algorithm which can iteratively minimize the function, and in that process, the algorithm

adjusts the weights up or down. RNN is facing two significant challenges that are exploding gradients and vanishing gradients.

Exploding gradients: Exploding gradients occur when the algorithm starts assigning unreasonable high significance to the weights. Luckily, the problem can be resolved by truncating or squashing the error gradients.

Vanishing gradients: The activation functions squish a more extensive input range into a smaller ranger. So the large change in the input of the activation function creates a slight change in the output; in turn, the derivatives produced become small. When the gradients are too small, the derivatives of each layer are multiplied down the whole network final to initial during backpropagation. So the gradients get decreased exponentially as propagation is done. Due to these effects, gradients become so small that the weights and biases cannot be updated effectively, and eventually, the model stops learning.

3.2.2 LSTM Neural Network (Why LSTM is better?)

A recurrent neural network (RNN) is typically applied to work on sequential data. The RNN learns the interdependencies between the sequential data by the information obtained from the succeeding time steps by using the recurrent connection (network loops) and by the concept of backpropagation through time during the learning process. The activation of the earlier time step is given as input to the current or recent time step for predicting motive in these loops. RNN is proven to handle short-term interdependencies; however, it does not perform effectively with long-term dependencies or longer sequences due to the gradient vanishing or exploding problem during the training [129,130], as discussed above.

The long short-term memory (LSTM) network is one of the particular types of the Recursive Neural Network(RNN), appropriate for processing and forecasting significant events with comparatively longer intervals and delays in the time series sequence data. LSTM is dissimilar compared to the RNN mainly because it uses a processor to the existing structure to decide if the information is valuable or not valuable.

The underlying structure of the processor is called a cell. It contains four interacting layers of networks activated by either the sigmoid function (σ) or the tanh function, all with their own different set of parameters. Each of these networks, also refered to as gates, have a different purpose. They will transform the cell state for time step t $c_{t,}$ with the relevant information that should be passed to the next time step.

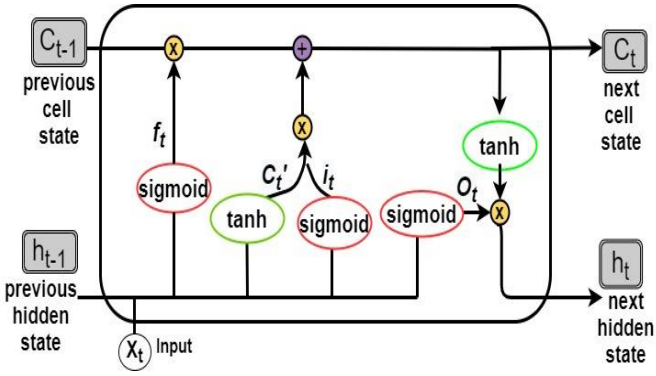
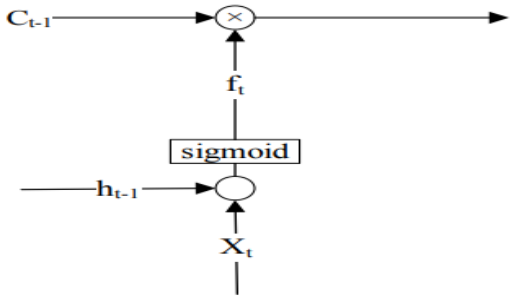


Figure 3.7. Internal gated structure of basic LSTM unit/cell

LSTM network was introduced to resolve the vanishing gradient issue of simple RNN and effectively operate the forecasting of longer input data sequences [131]. The conventional architecture of perceptron was replaced with the memory function to maintain the cell state and gated functions for controlling the input information.

The structure of LSTM includes three gates: an input gate, a forget gate, and an output gate. Figure 3.7 represents the architectural scheme of a single LSTM unit/cell. The internal cell state of the cell at time step *t* is represented as c_t, and at time *t-1* is symbolized as c_{t-1}. The inputs to the LSTM cell at time *t* are the input vector x_t, the hidden vector h_{t-1}, and the state of the cell c_{t-1}. LSTM generates the output c_t and h_t based on the given input and uses the three gates' functionality. The function of forget gate, input gate, and output gate f_t, i_t, and o_t respectively can be defined using equations 3.3 to equation 3.5.

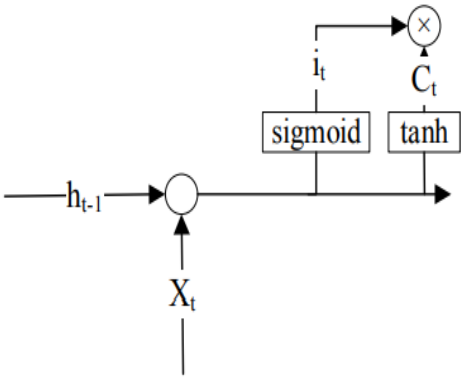
Forget gate:



The earlier hidden state h_{t-1} from the previous cell in the layer and the input given at current cell/unit x_t are provided to the sigmoid function. The output function f_t (between 0 to 1), calculated using equation 3.3, is multiplied with c_{t-1} to decide the degree of information to be forgotten at the forget gate.

$$f_t = \sigma(W_f\,[h_{t-1}, x_t] + b_f) \tag{3.3}$$

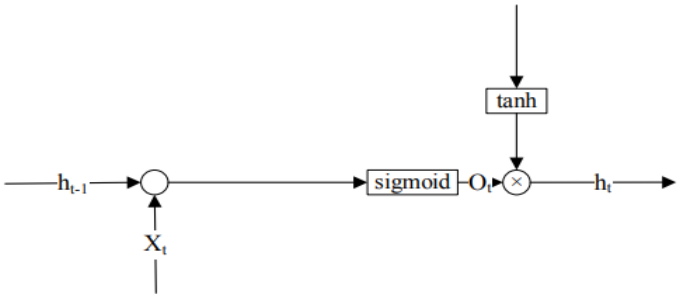
Input gate:



The sigmoid activation function at the input gate produces the value i_t in the range of 0 and 1 using equation 3.4, which determines the values to be updated or added in the neuron state. The intermediate cell state c_t' can be calculated using equation 3.6 by utilizing the previous time point outcome h_{t-1} and the current cell input x_t. The intermediary cell state c_t' and i_t are multiplied together element-wise, and then the result is added with the output of forget gate for calculation of the cell state c_t at time t as per equation 3.7.

$$i_t = \sigma(W_i\,[h_{t-1}, x_t] + b_i) \tag{3.4}$$

Output gate:



The output gate determines the information to be given as output in the state of the neuron. The output state of the input gate (c_t) is given to the *tanh* activation layer. The output from *tanh* between 1 and -1 is multiplied with o_t, calculated using equation 3.5 to produce the output hidden state h_t at timestep t as represented in equation 3.8.

$$o_t = \sigma(W_O\,[h_{t-1}, x_t] + b_O) \tag{3.5}$$

$$c_t' = \tanh(W_C\,[h_{t-1}, x_t] + b_C) \tag{3.6}$$

$$c_t = f_t * c_{t-1} + i_t * \ c_t{}' \tag{3.7}$$

$$h_t = o_t * \tanh(c_t) \tag{3.8}$$

Parameter *W* is weight, and *b* is the bias applied respectively at each stage. σ is the logistic sigmoid activation function, and *tanh* is the hyperbolic tangent function.