International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys.13, No.1 (Apr-23)

http://dx.doi.org/10.12785/ijcds/130157

Data Stream: Statistics, Challenges, Concept Drift Detector Methods, Applications and Datasets

Hussein A. A. Al-Khamees[1], **Nabeel Al-A'araji**[2] **and Eman S. Al-Shamery**[3]

[1,2,3]*Software department, Information Technology College, Babylon University, Babil, IRAQ*

Received 18 Feb. 2022, Revised 20 Dec. 2022, Accepted 6 Feb. 2023, Published 16 Apr. 2023

Abstract: Through real-world applications, the data stream is generated. In contrast to traditional data, data streams have different characteristics; a huge and endless amount, on-line arriving with high speed, single processing as well as the nature of being not static, in the sense that it evolves over time; this is the concept drift. Accordingly, the mining and analysis of the data stream is an arduous and attractive task. Various frameworks for data stream (analysis, mining, etc.) have been proposed over the past years. In the same context, identifying the number of classes of data streams is an important initial step when designing a model for processing the data stream. At present, deep neural networks (DNNs) are a fundamental technique in various applications. DNNs have many structures, including the multilayer perceptron (MLP). In this paper, we propose a deep neural network (DNN) model based on multilayer perceptron (MLP) to classify the streaming datasets and detect their classes as an analysis step for this data type. The proposed model tests different synthetic and real-world stream datasets. The results proved that this model detects the actual number of classes for the given stream dataset. Moreover, this paper presents a systematic review of data stream, its statistics, challenges, concept drift detector methods, data stream applications in different sectors in addition to the streaming datasets.

Keywords: Data Stream, Data Stream Challenges, Concept Drift Detector Methods, Streaming Datasets, Deep Neural Network, Multilayer Perceptron.

1. Introduction

Currently, a lot of people can share data anywhere and at any time easily. Through diverse applications in an unlimited field, a huge amount of data can be generated. This data type is known as ***Data Stream***; that can be defined as a massive amount of data which is arrived online with a high speed and ordered sequence, not static but evolving over time, concept drift appears due to the change of data distribution and mostly has a high dimensional [1].

Many sources are able to generate huge amounts of data in different structures, where the data stream is one of them. Some of these sources are sensors, satellite data, surveillance and tracking, network traffic, stock market, retail and so on [2].

In reality, the characteristics of data stream differ from their counterparts in traditional data, and this is due to the unique ones of this data. Usually, in traditional data, the data is static, and the most important aspect is that the data is available where it is stored in advance; therefore, it is described as limited data [3]. In more detail, the main differences between the data stream and traditional data are described in [1].

Data stream models aspire to provide analysis and knowledge extraction from a boundless amount of data to provide real-world data processing that was not previously described. Therefore, like traditional data, many techniques can be implemented for the data stream such as classification, clustering, frequent patterns and regression [4], [5].

The most important feature of the data stream is its environment is non-stationary; so, its behaviour is evolving over time. This leads to a concept called ***The Concept Drift***. Therefore, the classifier must deal with this challenge carefully to detect and adopt it if appears since it causes a decrease in the overall accuracy. As a result, the assumption of the persistence in the data stream environment (or its distribution) is incorrect [6].

Another significant aspect, represented by data learning (training) as the training methods of the traditional data cannot in any way be implemented on the data stream. Obviously, the pre-processing steps are still the same for both. The adaptive learning method can be implemented for detecting the concept drifts in streaming data perfectly [7].

As a result, because of these challenges, the traditional

E-mail address: hussein.alkhamees7@gmail.com, nhkaghed@itnet.uobabylon.edu.iq, emanalshamery@itnet.uobabylon.edu.iq

methods are unsuitable to address and analyze the data stream. Various frameworks for data stream mining have been proposed and developed over the past years. Therefore, it has become an interest center for researchers [8].

Deep learning is a machine learning technique depended on the deep neural networks (DNNs), which proved a great success in different applications in recent years. However, DNNs have many architectures, including the multilayer perceptron (MLP) and also many algorithms to train it, the most used is the back-propagation (BP) algorithm (Bahrain or Springer paper).

Our goals in this paper are to present:

● A deep neural network (DNN) model based on multilayer perceptron (MLP) architecture and the back-propagation (BP) algorithm to train this network. The main intention of this model is to classify the data stream and then detect their classes as an analysis step to the data stream. This (DNN) model tests both the synthetic and the real-world stream datasets. The results proved that the proposed (DNN) model detects the actual number of classes for the given stream dataset.

● A systematic review of the data stream, its statistics, challenges, concept drift detector methods, data stream applications and stream datasets. This is needful to enrich the researcher with (1) the data stream and most of its aspects; (2) determine its implementation areas; and (3) explain and discuss its challenges.

The rest of the paper is organized as follows: Sec. 2 discusses the related works. Sec. 3 explains the data stream statistics while Sec.4 displays the data stream aspects. The data stream challenges are illustrated in Sec.5. Sec. 6 demonstrates the concept drift and the concept drift detector is summarized in Sec. 7. The applications of the data stream are presented in Sec. 8. The proposed DNN model and its techniques are explained in Sec. 9. The streaming datasets are displayed in Sec. 10. Finally, Sec. 11 illustrates the conclusion of the paper.

2. RELATED WORKS

Few studies have dealt with the data stream in its precise detail, and we may not find a paper that collects a description of the data stream, its facts as statistics, challenges and applications. Therefore, this paper was presented to be comprehensive to describe the data stream, its statistics, its challenges and its applications.

In [9], firstly the authors discussed the tasks that can be implemented on the stream mining, and how those tasks play a decisive role in real-world. These tasks are classification, clustering, frequent patterns and regression. The paper displayed the data stream challenges and their processing types. Then each of the above techniques is explained, and finally, the evaluation metrics in the stream are illustrated.

According to [10], the paper clarified the processing algorithms of the data stream as well as, the method to extract knowledge from data. The authors have summarized the pre-processing techniques such as data cleansing, data integration and data transforming, while the data reduction aims to decrease the dataset size. Finally, the authors classified the solutions of treating data stream into data-based and task-based solutions. The solutions for data-based contain aggregation, data sampling, sketching, load shedding, synopsis structures and wavelets. While the solutions of task-based consist of approximation algorithms and time windows.

The authors in [11] classified the computing of big data into two kinds according to the processing requirements, batch computing and data stream (real-time) computing, then they explained the differences between them in terms of processing. The main issues in the data stream from the authors' point of view are scalability, integration, fault-tolerance, timeliness, consistency, load balancing, high throughput, privacy, accuracy. Finally, the authors displayed the techniques for data stream analysis, such as clustering, classification, fuzzy techniques.

3. DATA STREAM STATISTICS

Roughly, two-thirds of the world's population will have access to the Internet in 2023. Totally, Internet users will amount to 5.3 billion (66 % of the world's population) in 2023, while in 2018, Internet users around the world were 3.9 billion (51 % of the world's population)[1].

While the digital world contained approximately 2.7 zettabytes, or 2.7 trillion gigabytes of data for the first half of 2020 [12].

Data stream specialists expect the amounts of data generated to continue to rise dramatically in the coming years to reach unprecedented rates (180 zettabytes by 2025) [13]. Below are some examples to imagine the huge amount of data stream for some well-known sites around the world that generated data stream [1], [14]:

● The search engine Google, handles 40,000 search queries each second, this means; around 3.5 billion in a day and 1.2 trillion search queries yearly. Moreover, the voice search in Google in 2016 increased 35 times when compared to its counterpart in 2008.

● National Aeronautics and Space Administration (NASA) produce about 4 TB images in a day.

● Walmart (USA retail company) treats around 20 million transactions in a day.

Social networking sites represent one of the main sources of data stream through which, many effective models have been built to address a specific problem. The most

[1] https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html

important platforms of these social networking sites are Facebook, Twitter and Instagram as follows[2]:

● Facebook platform generated a data rate of about 4 petabytes per day in 2020.

● Twitter platform users generate nearly 500 million tweets per day.

● Daily, 95 million videos and photos are shared in Instagram platform.

● There are 306.4 billion emails sent around the world every day.

4. DATA STREAM ASPECTS (V5)

Besides the above challenges, the data stream models must cope with different aspects of the data stream. Sometimes these aspects are known as dimensions. Therefore, there is a need to deal with Volume, Velocity, Variety, Veracity and Value which are mostly known as V5 [15].

● **Volume**: indicates the massive amount of data that is generated in real-time.

● **Velocity**: refers to the fast data generation (their rate of growth) or how quickly data is collected to be analyzed. Therefore, researchers must use a minimum amount of time.

● **Variety**: is related to the type of data structure like structured, unstructured and semi-structured. On the other hand, an appropriate algorithm/s should be used and most likely an advanced one.

● **Veracity**: refers to both availability and accountability.

● **Value**: refers to how to get the higher level of knowledge through the using of the mining techniques implemented on the data stream.

At the beginning of dealing with data streams, its models were described as three-dimensional (3D) that included (Volume, Velocity, Veracity). After that, a fourth dimension was added, which is Veracity. And then the fifth dimension that is Value was added. These five dimensions form the concept of V5 [16].

5. DATA STREAM CHALLENGES

The data stream has different characteristics from traditional data and this may due to the new challenges in the data stream. This section will discuss the most important challenges extracted from the data stream, which are as follows:

1) Massive amount of data: while the data stream is a huge amount of data (endless arriving of data), where their samples arrive continuously and sequentially [17]. This vital aspect leads to generate a number of challenges to the data stream:
● Bounded memory (limited storage size).
● The whole data can't be stored, hence, techniques of the data reduction are implemented such as windowing, sampling, synopsis, . . . , etc.
● In the processing step, a single pass is allowed to every sample, then, it can be removed or stored if needed (because of the high costs of storage devices).

2) Fast online access: the data stream sources are able to generate the data stream consistently and send its data speedily. Moreover, these data samples are appearing sequentially (over time) and the method of arriving data cannot control. Accordingly, the challenge of data stream is:
● Data stream samples are on-line arriving at high speed from sources.

3) Data stream techniques: many techniques for data stream can be implemented depending on the type of problem to be solved and the algorithms applied in the design model. Most of them are linear and sub-linear in time complexity of the algorithms that usually implemented in real time. Accordingly, the challenges for data streams are:
● The response of the model should be immediately (providing the user with results at any time it requests). Simply, it's promptly real-time data analytics.
● Running time is very crucial since data samples must be processed as quickly as possible, otherwise the algorithm will be ineffective.

4) The data stream has unexpected properties besides, in general, it classifies as heterogeneous data. This state of heterogeneity resulted from the heterogeneous sources. Furthermore, as long as there is diversity in the data stream applications implemented in a variety of disciplines, the system encounters uncertain data and also the presence of a specific class label more than the other. Therefore, the challenges are: :
● High dimensional data.
● Imbalanced labels of the data samples.

5) The nature of the data stream: unlike traditional data, where the data is static, the data stream is dynamic that is evolving over time. This may be due to the non-stationary of the data stream environment. The interesting issue is that this dynamism and non-stationary is causing one of the most important data stream challenges and the most studied, analyzed and designed models to detect, that is the concept drift [18]. The concept drift occurs because of the changing of the data distribution over time. This appearance causes the performance of the model to deteriorate, and the model's results will be less accurate (or inaccurate). Wherefore the algorithm must be modified (updated) to deal with any changes that occur in the data. Hence,

[2] https://techjury.net/blog/how-much-data-is-created-every-day/gref

the model is automatically updated for adapting the most recent changes. Thus, the challenge is [19]:

- ***Concept drift.***

6. CONCEPT DRIFT

It is one of the major challenges of data stream and is considered one of its most important areas. Many of the previous works focused on the concept drift, its nature, causes of its occurrence, its types, and most importantly, how the system is trained to detect it and methods of detection. Certainly, the system must detect it when it appears and adopt this change in behavior and the structure of the system. In other words, the system can't remain constant in a non-stationary environment. The reason for this great and important interest is that the concept drift eventually causes the results to be inaccurate [2].

In terms of name, the concept drift consists of two parts the first is a concept that refers to the distribution of the data at a specific time. While the second is drift, that means the parameters are changed [20].

A. The Detection of Concept Drift

Depending on the aspect of input-target domains, the concept drift can be defined as the change in the output domain between two times (t_0 and t_1) as can be shown in Fig. 1.

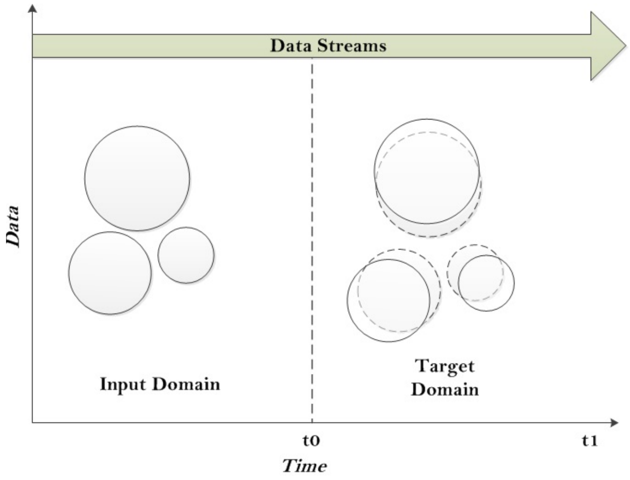


Figure 1. The appearance of concept drift

The rapid detection and adoption of concept drift represent an additional and important feature of the system that handles data streams because it causes the rising of the accuracy of the final results [21].

Certainly, one of the systems' strengths of learning from the data stream in non-stationary environments that evolve over time is the ability to detect and adopt the concept drift [19].

The general idea of the data stream systems is to receive data from external data sources, often by applications or devices. Then, the system plays a very important and effective role, which is similar to a control center for that data, its purpose is to extract useful data from the huge amount of receiving data. In other words, the system is able to process this data through the implemented techniques. After that, the system is evaluated by metrics. Finally, the system is ready to use and can receive the new data to achieve the goal for which it was designed.

B. Examples of Concept Drift

This section covers some examples of concept drift [6], [22]:

- In Network Intrusion Detection systems (NIDs), while the streaming data consists of the normal cases, an attacker is looking for new methods to get around security systems and then logging and accessing as a normal case.

- In the system of following personal interests, a user changes its interests (preferences) according to a new personal conviction or acquisition of new knowledge or reaction to a certain event and so on.

- In the medical decision aiding systems, stream data refers to the normal medical conditions of the patient and then the emergence of a change in the data according to a specific reason, such as the response to applied drugs (whether it is a positive or negative response) and changes in the patients' resistance and many other reasons.

- In the sensors, there is a problem that cannot be ignored, which is represented at the end of its operational life (reach the stage of un-reading data) or what is called in the industrial world aging. Sometimes, sensors suffer from malfunctions in their electronic parts. These reasons lead to a change in the reading of the data.

- In predicting systems for electric power consumption, the data stream is stationary until there is a change occurs in this data, such as overpopulation that leads to the construction of new cities or neighborhoods, and thus a change (increase) in demand for consumption, improvements in energy production efficiency. As well as weathering fluctuations that get throughout the year.

- In the wind turbine systems, there are several reasons for the appearance of the concept drift. However, these reasons can mainly be classified into two types: atmospheric factors (such as wind speed, temperature,...,etc) and mechanical factors (such as aging, maintenance, failure,...,etc).

C. Mathematical Representation of Concept Drift

Mathematically, suppose data streams (DS) represented by infinite sequences of samples as: DS $=\{S_1,S_2,\ldots, S_i\}$ where, i refers to the total number of samples until yet, and any S_i is arrived at a specific time, and generated from a distribution D_j. In the case of the data stream, the distributions are different such as S_i is from D_j whilst S_{i+1} from D_{j+1}.

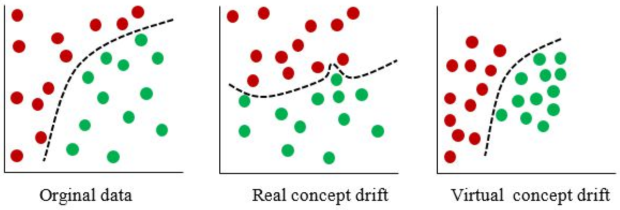
In contrast, in a stationary environment, all data samples S_i are generated from the same distribution D_j.

Regarding the concept drift, let's represented it by (P), it occurs when two samples are arrived at two different times (t_0) and (t_1) and their distributions are changed. However, it can be expressed by equation (1) [7]:

$$S_{t0}(D_j) \neq S_{t1}(D_{j+1}) \tag{1}$$

D. Concept Drift Types

Basically, the concept drift is either real drift or virtual drift [20]. The real drift is also known as class drift and prior-probability shift [8]. While the virtual drift is known as covariate drift [23]. A third type can be added that is the population drift [8]. Fig. 2. shows the differences between original data, real and virtual concept drifts.



Orginal data Real concept drift Virtual concept drift

Figure 2. The difference between original data, real and virtual concept drifts

In fact, there is no definitive way to classify the concept drift. Thus there is a diversity in the classification methods presented by researchers. Practically, the concept drift has been classified according to [24]:

- The speed of change.
- The change distribution.
- The severity.

The change speed (sometimes known as shift pace) can measure through the total number of time steps required for completing the drift. This time step refers to arrive one (or more) data samples during a fixed time. Usually, the smaller time to complete the concept, the faster it will be [25]. Depending on the change speed, the concept drift is classified into [2], [24]:

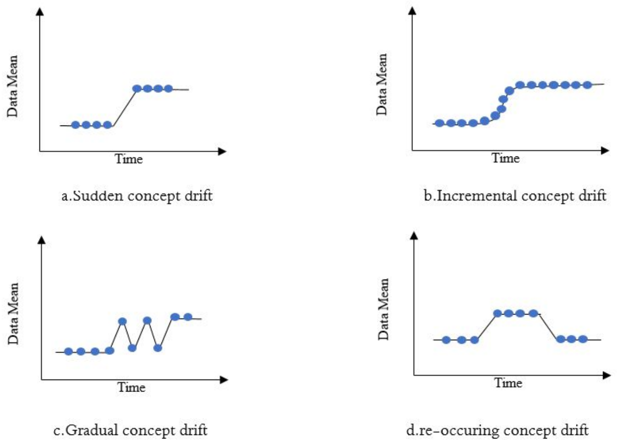
- **Sudden or Abrupt drift**: when the drift can complete in only one time step. Mathematically, in this type, the drift is represented according to: S_{i+1}, but $D_j \neq D_{j+1}$.

- **Incremental drift**: when the drift is completed in more than one time step. In this type, there are some samples represent intermediate cases between the completeness stages. In the incremental drift, the speed of change is an unimportant criterion, where the difference between D_j, D_{j+1} is not very important.

- **Gradual drift**: is similar to incremental drift where it needs more than one time step to complete. This type happened if the concept drift appears in a gradual mode. Mathematically, in this type, the data samples of S_{i+1} have been generated from both distributions D_j and D_{j+1}.

- **Re-occurring drift**: if an old concept drift re-appears, this type is known as re-occurring drift.

These four types of the concept drift are illustrated in Fig. 3.



a.Sudden concept drift b.Incremental concept drift c.Gradual concept drift d.re-occuring concept drift

Figure 3. The types of concept drift where every dot indicates to a time step

In addition to the concept drift, the data stream has other changes, such as concept evolution and concept forgetting. Sometimes, the system needs to forget some data samples that do not correspond to the general nature of the data. The reason for the correlation in the first two types is that the model always needs to deal with the most recent data samples. While the concept forgetting gives up the old data samples [26].

7. CONCEPT DRIFT DETECTOR

Many methods have been suggested over the past years that handle the data stream generally and concept drift especially. In those methods, several classifiers are used to predict the data labels and many of them involved the step of concept drift detection; or the so-called concept drift detector models. Therefore, many researchers that designed those models take into account the detection of concept drifts and then adopt it [27].

A. Designing the Models

The general idea of the detector models is either tracking the classifier error or computing the statistical characteristics of the dataset.

Every model (detector) has specific parameters which differ from one model to another, these parameters also have different values. Basically, the parameters and their values are matching the characteristics and behavior of the dataset to achieve the desired goal of the model. As long as the concept drift is within the behavior of the data, it is natural for the model to deal with these concepts and then rapidly detect and adopt them by updating their structure [28].

Updating the model is an important challenge to maintain the effectivity of the model over time, besides some other aspects must be preserved such as accuracy, efficiency and ease of use [27], [29].

Depending on [8], a model should have many requirements to be effective in dealing with streaming data through the non-stationary environments, which are:

- Must distinguish noise from drift to be not confused with models that operate in stationary environments (with static data).
- Able to detect a concept drift in a little time.
- The step of processing is done in a small period of time.

According to [30], several indicators can be observed when concept drifts occur, such as performance measures, classifier model characteristics and data characteristics.

B. Concept Drift Detector Methods

Many methods have been developed to deal with the concept drifts that can be classified into four major categories which are [8]:

1) **Sequential analysis-based methods**: the main idea is to send a specific alarm if a determined threshold exceeds.
2) **Statistical-based methods**: these methods are based on the computations of some statistical parameters such as the standard deviation, mean, variance, . . . , etc.
3) **Methods based the change distributions on two various time windows**: these methods depend on applying a predetermined window and then summarizing the previous information, lastly, a sliding detection window is used for the most recent or current examples.
4) **Contextual-based methods**: these methods depended on using a timestamp of the data samples as input features to the batch classifier.

The methods which are based on the change distributions on two various time windows are widely used, especially during the last few years. Many models belong to this method such as ADWIN, CVFDT, FHDDM, HDDM-A, HDDM-W test, SeqDrift2, KSSVD and MDDMs [31].

C. Examples of Concept Drift Detector Models

Before the start of the new century, a number of detector models were proposed such as STAGGER (1986), FLORA (1996). Then, at the beginning of the new millennium, Very Fast Decision Tree learner VFDT (2000), Concept-adapting Very Fast Decision Tree learner CVFDT (2001), Streaming Ensemble Algorithm SEA (2001), Accuracy Weighted Ensemble AWE (2003) [28], [29].

Thereafter, a vast number of concept drift detector models have been presented including: Drift Detection Method (DDM), Early Drift Detection Method (EDDM), Statistical Test of Equal Proportions (STEPD), Paired Learners (PL), ADaptive sliding WINdow (ADWIN), Exponentially Weighted Moving Average (EWMA), EWMA for Concept Drift Detection (ECDD), Fast Hoeffding Drift Detection Method (FHDDM), Hoeffding Drift Detection Method with the A-test (HDDM-A), Hoeffding Drift Detection Method with the W-test (HDDM-W), Fisher Test Drift Detector (FTDD), Wilcoxon Rank Sum Test Drift Detector (WSTD), Sequential Drift (SeqDrift), Sequential Drift2 (SeqDrift2), McDiarmid Drift Detection Method (MDDM), KS-SVD test, Reactive Drift Detection Method (RDDM), Drift Detection Methods based on Hoeffding's Bounds (HDDM), Page Hinkley Test (PHT), SEED Drift Detector (SEED) and Equal Means Z-Test Drift Detector (EMZD) [8], [27].

8. DATA STREAM APPLICATIONS

The applications of data stream have been widely used over the past years. After studying and analyzing more than 40 scientific papers in this area, we have summarized the data stream applications in five main fields which are computer, medicine, environment, economic and miscellaneous fields. It is in fact endless fields and applications [1], [2], [12], [14], [17], [19], [20], [22], [32], [33]:

- **Computer field**: such as social networks, website analyses, web click streams, analysis of the Internet log, spam filtering, Internet of Things (IoT), robots, ubiquitous computing, cyber-attacks monitoring systems, Network Intrusion Detection systems (NIDs), prediction of network load, sensor networks, sensors to monitor border security, surveillance systems, smart systems, smart cities, smart home, prediction of traffic congestion in smart cities, and others.

- **Medicine field**: for example, healthcare sensors, patient tracking sensors, disease prediction systems, disease diagnosis, medical science data, medical decision aiding, electronic medical records, health technologies systems, the systems of diagnosis, prevention, monitoring and management of chronic diseases by wearable devices/sensors, prediction of heart beat irregularities in medicine and so on.

- **Environment field**: such as prediction of the natural disasters, weather monitoring, detecting changes in weather temperature, detecting changes in water temperature, air traffic control and navigation, car navigation systems, earthquake forecasting systems, prediction of floods, sensors systems for the temperature, humidity and light, vibration, pressure, monitoring the astronomical data, meteorological analysis systems, climate data analysis, monitoring of air pollution, prediction of sun spot activity, wind turbine systems and others.

● **Economic field**: for example, financial transactions, identification of customer preferences, analysis of the stock market, financial data prediction, credit card fraud detection, financial fraud detection, customer profiling, sales prediction, banking services systems, loan systems, insurance systems, debit card systems, prediction of future sales and so on.

● **Miscellaneous field**: such as online shopping recommendation systems, telecommunication systems, GPS device tracking, mobile device tracking, phone records systems, electric power consumption systems, predicting electric power consumption systems, news group filtering, monitor water distribution networks, monitor gas distribution networks, astronomy systems, sentiment analysis, system following personal interests, emergency response systems, recommendation systems and others.

9. THE PROPOSED MODEL

Our proposed DNN model aims to classify the data stream and detect their classes by using multilayer perceptron (MLP) architecture and back-propagation (BP) algorithm to train this network. This section explores the major techniques that are used in the proposed model.

A. Machine Learning

Machine learning is the spine of the artificial intelligence (AI). From the data being processed, the machine learning models are learning and building their structure. The most prominent characteristics of ML models are the representation and adaptation. Within this realm, the machine learning is always presenting new methods to analyze the data stream [34],[35].

B. Deep Neural Networks (DNNs)

The structure of a neural network consists of three layers: input, hidden layer, and output layer. The neural networks are either shallow or deep. The major difference between them is that in the shallow neural networks, there is a hidden layer whilst, in the deep neural networks, there are two or more hidden layers. However, this number varies from one DNN model to another [36].

Deep learning that based on the deep neural networks (DNNs) is the most efficient and effective technique among all the machine learning techniques [37].

C. DNN Structure

Each layer consists of neurons (sometimes known as artificial neurons) that differ in number from one layer to another [36].

In a neural network, each neuron is connected to other neurons in adjacent layers and this connection is represented by the weights [34].

DNNs have different architectures, the most common and successful one is the multilayer perceptron (MLP). Simply, it is defined as a feed-forward neural network with multi hidden layers.

D. MLP Training

Various algorithms for MLP architecture training have been proposed and developed previously. Nevertheless, the back-propagation (BP) algorithm is the most common and effective algorithm for training the DDNs [36].

The back-propagation algorithm was adopted in the proposed model for its ability to continuously update and optimize the network parameters, especially the network weights based on the training dataset [34].

Fig. 4 shows the proposed model framework.

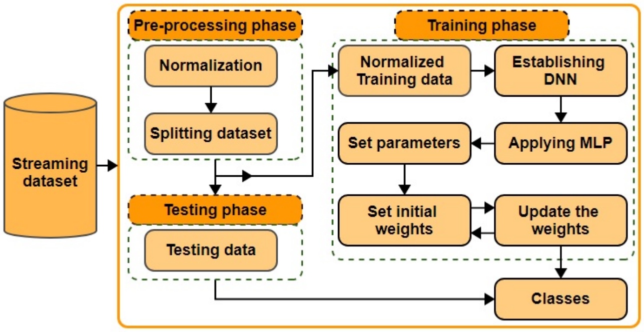


Figure 4. The general framework of the proposed model

As can be seen from Fig. 4, the proposed (DNN) model consists of three phases as follow:

- **Pre-processing phase:** it contains two steps:

1) **Normalization:** Normalization techniques transform the input values of the dataset in such a way as to ensure that they will be in the same scale (range) thus it is reflected in improved performance and network stability [38]. The normalization techniques include, Min-Max, decimal scaling, Z-score and others. In this model, we used the Min-Max normalization technique. In general, the speed of machine learning models is increased when using the normalization techniques.
2) **Splitting datasets**: The dataset used by a machine learning model is usually divided into two parts: a training part (training data) and a testing part (testing data). Over the past years, many methods and techniques have been proposed to split the dataset. The cross-validation technique is the most common one which is used to split the dataset into training data and testing data [39]. However, in this model, we applied this technique with the ratio, 70 % as training data and 30 % as testing data.

- **Training phase:** it consists of the following steps:

1) **Normalized Training data:** Firstly, this phase uses the normalize training data that was split in the previous phase.
2) **Establishing DNN:** Then, we establish the deep neural

network.

3) **Applying MLP:** After that, we apply the multilayer perceptron (MLP). architecture.
4) **Set parameters:** In this step, we initialize the network parameters such as the number of hidden layers is three, the number of neurons in each hidden layer as follows: 100, 75, and 50 respectively, activation function is ReLU, the initial learning rate is 0.001.
5) **Set initial weights:** This step is responsible to set the initial weights to start the training process in the proposed model.
6) **Update the weights:** This step for updating of the weights. The back-propagation algorithm continuously update and optimize the weights until obtaining the optimal weights.

- **Testing phase:** It's the last phase in the proposed model, the testing dataset from the pre-processing phase is used which forms 30 % of the total streaming dataset. This dataset applies to the classes that represent the results of the training phase.

Finally, the results of the proposed deep neural network (DNN) model are the classes which are actually matching the classes in the streaming dataset. However, these results will be shown in the next section.

10. STREAMING DATASETS

Generally, the stream datasets that are used to evaluate the systems can be classified into two types, synthetic datasets and real-world data sets [40].

- **Synthetic Stream Datasets**: Sometimes known as the artificial dataset. Based on many available datasets and after studying and conducting the appropriate analysis for it, the synthetic stream datasets are generated by the researchers. So, many studies that contain this type of dataset are known by local names, for example DS1, which indicates the Dataset number 1 and so on.

The synthetic datasets have a number of properties distinguished from real-world datasets, such as [41], [42]:

1) The testing of hypotheses is easier to implement, these tests may include the high dimensionality scaling and also the robustness of noise.
2) In the synthetic datasets, it is easy to control in:
   a) The data samples number.
   b) The clusters number.
   c) The dimensionality.
   d) Various characteristics of both distributions or evolution.
3) It can be re-produced in easy methods.
4) Inexpensive in terms of storage or transportation.
5) It can easily simulate the required drift style.

We analysed six synthetic streaming datasets, these datasets are illustrated in Table I, where # Classes is the number of classes that are detected by the proposed model.

TABLE I. CLASSES NUMBER OF SYNTHETIC STREAMING DATASETS THAT ARE DETECTED BY THE PROPOSED DNN MODEL

Seq.	Stream Dataset Name	# Classes
1	STAGGER	2
2	Mushroom	2
3	LED	10
4	Waveform	3
5	SEA	2
6	Rotating Hyperplane	2

More details of these synthetic streaming datasets, as well as the learning task are displayed in Table II [14], [42], where # Actual Classes is the number of actual classes and # Att. is the number of attributes in the stream dataset.

- **Real-World Stream Datasets**: The second type of the streaming dataset is the real-world. Generally, a smaller number of this type is available to the public and free downloaded. In this paper, fifteen real-world stream datasets were analyzed and tested by the proposed model. More descriptions of these real-world stream datasets are as follows:

1) **Sensor Stream**: it contains data of (time, temperature, humidity, light, and voltage) which are gathered from 54 sensors for 2 months, where a single reading each 1-3 minutes. The description of this dataset is as follows: [2,219,803 samples, 5 features, and 54 classes].
2) **Powersupply Stream**: this data set recorded the power supply of the electricity company in Italy that collected their samples from two sources hourly through the years 1995 to 1998. The powersupply description is: [29,928 samples, 2 features, and 24 classes].
3) **Hyper Plane Stream**: sometimes called HyperP stream, however, it's the only synthetic data stream among the other datasets inside a website. This data set involves gradually evolving (concept drift) defined by many equations. Its description can be summarized as:
   [100,000 samples, 10 features, and 5 classes].
4) **Kddcup99 stream**: it is derived from the KDD CUP – 1999, which has been implemented by many models in order to separate intrusions connections from normal ones. Some researchers classified this data set as not real-world dataset, due to the shortage of the public datasets for Intrusion Detection System (IDS) [43]. This dataset contains:
   [494,021 samples, 41 features, and 23 classes].
   Note that the original Network intrusion detection consists of 4,898,431 samples with the same numbers of features and classes.
   However, the first four datasets can be download free[3].

[3] https://www.cse.fau.edu/~xqzhu/stream.html

TABLE II. DESCRIPTION OF SYNTHETIC STREAMING DATASETS

Seq	Stream Dataset Name	Year	# Att.	# Actual Classes	Learning Task
1	STAGGER	1986	3	2	The learning task of the STAGGER streaming dataset is to classify the current sample into either a positive or negative class.
2	Mushroom	1987	22	2	The learning task of this streaming dataset is to predict the sample health, whether it is an edible and poisonous.
3	LED	1984	24	10	The number of classes is ten which has been reduced to only two. However, the learning task of the LED streaming dataset is to predict the current signal of the LED, on or off.
4	Waveform	1984	40	3	The Waveform learning task is to distinguish three different classes (0,1, and 2) of waveform.
5	SEA	2001	3	2	The learning task of the SEA streaming dataset is to predict the belonging of the sample to either class1 or class2.
6	Rotating Hyperplane	2001	10	2	Rotating Hyperplane learning task is to predict the class of the current instance, if its class is a positive or a negative class.

5) **Poker-Hand**: which was presented in 2002 [44]. As long as the pocket game consists of 52 decks, then every sample of this dataset relates to an example of the hand's content of 5 cards. Taking into account the cards order is an important aspect in this dataset. Its details are as follows:
[1,025,010 samples, 11 features and 10 classes].

6) **Electricity**: it has been gathered from the Market of Electricity in New South Wales, Australia, therefore, it is affected by demand and supply. The prices are constantly changing, and they should be determined every five minutes. The dataset collected from 1996-1998 and each sample was created during 30 minutes, so it produced 48 samples in a day. However, the electricity dataset consists of:
[45,312 samples, 8 features and 2 classes].

7) **SpamAssassin**: it presented in 2009 [45]. Spam can define as any unwanted email, which has become one of the biggest problems on the Internet. The description of the Spamassassin as follows [46]:
[9,324 samples, 97,851 features and 2 classes].

8) **NOAA**: it was proposed in 2011 [47], it is a weather dataset collected from hundreds of stations over the world by the National Oceanic and Atmospheric Administration (NOAA). Many measurements were recorded daily like speed of the wind, pressure degree and the visibility level. This dataset contains:
[18,159 samples, 8 features and 2 classes].

9) **Forest Covertype**: it was proposed in 1999 [48]. In this dataset, all the digital spatial data have been gotten from:
1- US Geological Survey (USGS).
2- US Forest Service (USFS).
This dataset was collected from 4 wilderness areas representing a system of cells, each cell refers to an area (30 x 30 meters). This dataset has:
[581,012 samples, 54 features and 7 classes].

10) **Gas Sensor Array**: it was proposed in 2012 [49]. This dataset was collected from 2007-2011 at the California University, San Diego, USA. It has 13910 samples of the time series sequences. The idea behind Gas Sensor dataset is to measure sixteen sensors of chemical gas stored in an array. The Gas Sensor Array dataset consists of:
[13,910 samples, 128 features and 6 classes].

11) **Outdoor Objects**: it was suggested in 2015 [50]. The Pioneer platform makes use of a camera to capture several objects (40 ones) on the ground. Then, the next step is to determine which object is found in a picture. This dataset contains:
[4,000 samples, 21 features and 40 classes].
The datasets (5, 6, 7, 8, 9, 10 and 11) can be downloaded free[4].

12) **Credit Card Fraud Detection**: it includes information made on European credit cards in just two days of September 2013. This dataset contains 284,315 as normal transaction cases while 492 as fraud cases. So, this dataset was classified as an unbalanced dataset. However, its description is as follows:
[284,807 samples, 31 features and 2 classes].
The website to download this dataset is[5].

13) **NSL-KDD**: the main shortcoming in the KDD dataset, it has a massive number of redundant records, that led

[4] https://sites.google.com/view/uspdsrepository
[5] https://www.kaggle.com/mlg-ulb/creditcardfraud

to many algorithms to be biased towards those frequent records. Hence, it will have an effect on the final results of the model. Accordingly, to overcome the imperfections described above, an enhanced version of the original KDD'99 dataset was presented, known as the NSL-KDD dataset. In the new dataset, each redundant record is omitted besides, all the records are re-balanced. As a result, the NSL-KDD dataset becomes more realistic and practical for evaluating algorithms. NSL - KDD dataset has the description:
[148,517 samples, 41 features, and 5 classes].
However, this dataset can be downloaded free[6].

14) **Keystroke Dynamics**: it was proposed in 2009. The purpose of this dataset is to recognize the user as they type a password through the rhythm of writing. However, the researchers assigned 51 subjects (typists) from inside the university to write one password, so that each user typed the word 400 times. This process took place during 8 sessions, separated by at least one day. This means that the recurrence was 50 per session. Note that the password that all users wrote is the same password (.tie5Roanl) which is considered as a strong password since it consists of 10 characters that involve: a special character (.), lowercase letters (tieoanl), one number (5) and an uppercase letter (R). The keystroke dynamics dataset has:
[20,400 samples, 33 features and 51 classes].
More details and free downloading to this dataset in various formats can be found in[7].
Furthermore, another subset of keystroke dataset was presented in 2015. This subset dataset has the following description:
[1,600 samples, 10 features and 4 classes].
This subset is also free to be downloaded[8].

15) **HuGaDB**: it was presented in 2017 [51]. The main objective of this dataset is to activity recognition from six inertial sensor networks. HuGaDB dataset contains 12 behaviors actions which involved both static and dynamic activities. Furthermore, the main HuGaDB dataset consists of 364 files and all of them have the same number of features. HuGaDB dataset has the description:
[2,111,962 samples, 39 features, and 12 classes].
The website to download HuGaDB dataset is[9].

After applying the proposed (DNN) model to these real-world stream datasets, the results are explained in Table III.

11. CONCLUSION AND FUTURE WORK

Nowadays, with the great advances in technology, various applications in the real world are capable of generating enormous amounts of data stream that need immediate processing. Most of these applications relate to our daily life, such as health, economy, sensors, etc., while their importance is to extract knowledge from this huge amount of data amidst previously unfamiliar challenges. Therefore, there is a real need to study and analyze this data.

TABLE III. CLASSES NUMBER OF SYNTHETIC STREAMING DATASETS THAT ARE DETECTED BY THE PROPOSED DNN MODEL

Seq.	Stream Dataset Name	# Classes
1	Sensor	54
2	Powersupply	24
3	Hyper Plane	5
4	Kddcup99	23
5	Poker-Hand	10
6	Electricity	2
7	SpamAssassin	2
8	NOAA	2
9	Forest Covertype	7
10	Gas Sensor Array	6
11	Outdoor Objects	40
12	Credit Card Fraud Detection	2
13	NSL-KDD	5
14	Keystroke Dynamics	4
15	HuGaDB	12

Deep learning based on the deep neural network is a machine learning technique that has proven successful and efficient in various applications in several fields, especially in the data streams. Due to the data stream importance, this paper presents a deep neural network model depended on multilayer perceptron (MLP) architecture and the back-propagation (BP) algorithm to train this network. The model aims to classify the stream datasets. In summary, the resulting classes of this model matched the actual classes in the given stream dataset. Moreover, we present in this paper a systematic review of the data stream, its statistics, its challenges, concept drift detector methods, data stream applications in several areas besides the available streaming datasets.

In future work, the authors will present a method for concept drift detection based on developing data stream clustering algorithm as this concept is the most prominent challenge for the data stream.

Acknowledgment

We would like to acknowledge the University of Babylon for providing adequate support and sponsorship for this paper.

References

[1] H. A. A. Al-Khamees, N. Al-A'Araji, and E. S. Al-Shamery, "Survey: Clustering techniques of data stream," in *2021 1st Babylon International Conference on Information Technology and Science (BICITS)*. IEEE, 2021, pp. 113–119.

[2] M. Althabiti and M. Abdullah, "Classification of concept drift in evolving data stream," *Emerging Extended Reality Technologies for*

[6] https://www.unb.ca/cic/datasets/nsl.html

[7] http://www.cs.cmu.edu/ keystroke/

[8] https://www.sites.google.com/site/nonstationaryarchive/

[9] https://www.kaggle.com/romanchereshnev/hugadb-human-gait-database

Industry 4.0: Early Experiences with Conception, Design, Implementation, Evaluation and Deployment, p. 189, 2020.

[3] D. C. Sujatha and J. G. Jayanthi, "A survey on streaming data analytics: Research issues, algorithms, evaluation metrics, and platforms," in *Proceedings of International Conference on Big Data, Machine Learning and Applications: BigDML 2019*, vol. 180. Springer Nature, 2021, p. 101.

[4] H. A. A. Al-Khamees, N. Al-A'Araji, and E. Al-Shamery, "Data stream clustering using fuzzy-based evolving cauchy algorithm," *International Journal of Intelligent Engineering and Systems*, vol. 14, pp. 348–358, 2021.

[5] H. A. A. Al-Khamees, W. R. H Al-Jwaid, and E. S. Al-Shamery, "The impact of using convolutional neural networks in covid-19 tasks: A survey," *International Journal of Computing and Digital Systems*, vol. 11, no. 1, pp. 189–197, 2022.

[6] G. Ditzler, M. Roveri, C. Alippi, and R. Polikar, "Learning in nonstationary environments: A survey," *IEEE Computational Intelligence Magazine*, vol. 10, no. 4, pp. 12–25, 2015.

[7] S. Ramírez-Gallego, B. Krawczyk, S. García, M. Woźniak, and F. Herrera, "A survey on data preprocessing for data stream mining: Current status and future directions," *Neurocomputing*, vol. 239, pp. 39–57, 2017.

[8] J. Gama, I. Žliobaitė, A. Bifet, M. Pechenizkiy, and A. Bouchachia, "A survey on concept drift adaptation," *ACM computing surveys (CSUR)*, vol. 46, no. 4, pp. 1–37, 2014.

[9] M. Bahri, A. Bifet, J. Gama, H. M. Gomes, and S. Maniu, "Data stream analysis: Foundations, major tasks and tools," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 11, no. 3, p. e1405, 2021.

[10] D. S. Medeiros, H. N. C. Neto, M. A. Lopez, L. C. S. Magalhães, N. C. Fernandes, A. B. Vieira, E. F. Silva, and D. M. Mattos, "A survey on data analysis on large-scale wireless networks: online stream processing, trends, and challenges," *Journal of Internet Services and Applications*, vol. 11, no. 1, pp. 1–48, 2020.

[11] T. Kolajo, O. Daramola, and A. Adebiyi, "Big data stream analysis: a systematic literature review," *Journal of Big Data*, vol. 6, no. 1, pp. 1–30, 2019.

[12] S. Priya and R. A. Uthra, "Comprehensive analysis for class imbalance data with concept drift using ensemble based classification," *Journal of Ambient Intelligence and Humanized Computing*, vol. 12, no. 5, pp. 4943–4956, 2021.

[13] S. Mehta *et al.*, "Concept drift in streaming data classification: algorithms, platforms and issues," *Procedia computer science*, vol. 122, pp. 804–811, 2017.

[14] H.-L. Nguyen, Y.-K. Woon, and W.-K. Ng, "A survey on data stream clustering and classification," *Knowledge and information systems*, vol. 45, no. 3, pp. 535–569, 2015.

[15] A. Bifet and J. Read, "Ubiquitous artificial intelligence and dynamic data streams," in *Proceedings of the 12th ACM International Conference on Distributed and Event-based Systems*, 2018, pp. 1–6.

[16] J. Debattista, C. Lange, S. Scerri, and S. Auer, "Linked'big'data: towards a manifold increase in big data value and veracity," in *2015 IEEE/ACM 2nd International Symposium on Big Data Computing (BDC)*. IEEE, 2015, pp. 92–98.

[17] R. N. Gemaque, A. F. J. Costa, R. Giusti, and E. M. Dos Santos, "An overview of unsupervised drift detection methods," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 6, p. e1381, 2020.

[18] H. S. Yazdi, A. G. Bafghi *et al.*, "A drift aware adaptive method based on minimum uncertainty for anomaly detection in social networking," *Expert Systems with Applications*, vol. 162, p. 113881, 2020.

[19] S. Wares, J. Isaacs, and E. Elyan, "Data stream mining: methods and challenges for handling concept drift," *SN Applied Sciences*, vol. 1, no. 11, pp. 1–19, 2019.

[20] V. Souza, D. M. dos Reis, A. G. Maletzke, and G. E. Batista, "Challenges in benchmarking stream learning algorithms with real-world data," *Data Mining and Knowledge Discovery*, vol. 34, no. 6, pp. 1805–1858, 2020.

[21] S. Ancy and D. Paulraj, "Online learning model for handling different concept drifts using diverse ensemble classifiers on evolving data streams," *Cybernetics and Systems*, vol. 50, no. 7, pp. 579–608, 2019.

[22] B. Krawczyk, L. L. Minku, J. Gama, J. Stefanowski, and M. Woźniak, "Ensemble learning for data stream analysis: A survey," *Information Fusion*, vol. 37, pp. 132–156, 2017.

[23] G. I. Webb, R. Hyde, H. Cao, H. L. Nguyen, and F. Petitjean, "Characterizing concept drift," *Data Mining and Knowledge Discovery*, vol. 30, no. 4, pp. 964–994, 2016.

[24] H. Hu, M. Kantardzic, and T. S. Sethi, "No free lunch theorem for concept drift detection in streaming data classification: A review," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 10, no. 2, p. e1327, 2020.

[25] K. S. N. Prasad, A. S. Rao, and A. V. Ramana, "Ensemble framework for concept-drift detection in multidimensional streaming data," *International Journal of Computers and Applications*, pp. 1–8, 2020.

[26] L. Rutkowski, M. Jaworski, and P. Duda, *Stream data mining: algorithms and their probabilistic properties*. Springer, 2020.

[27] S. G. Santos, R. S. Barros, and P. M. Gonçalves Jr, "A differential evolution based method for tuning concept drift detectors in data streams," *Information Sciences*, vol. 485, pp. 376–393, 2019.

[28] Ł. Korycki and B. Krawczyk, "Adversarial concept drift detection under poisoning attacks for robust data stream mining," *arXiv preprint arXiv:2009.09497*, 2020.

[29] S. Kadam, "A survey on classification of concept drift with stream data," 2019.

[30] D. Brzezinski, L. L. Minku, T. Pewinski, J. Stefanowski, and A. Szumaczuk, "The impact of data difficulty factors on classification of imbalanced and concept drifting data streams," *Knowledge and Information Systems*, vol. 63, no. 6, pp. 1429–1469, 2021.

[31] O. A. Mahdi, E. Pardede, N. Ali, and J. Cao, "Diversity measure as a new drift detection method in data streaming," *Knowledge-Based Systems*, vol. 191, p. 105227, 2020.

[32] A. Zubaroğlu and V. Atalay, "Data stream clustering: a review," *Artificial Intelligence Review*, vol. 54, no. 2, pp. 1201–1236, 2021.

[33] M. Uddin and S. Syed-Abdul, "Data analytics and applications of the wearable sensors in healthcare: An overview," *Sensors*, vol. 20, no. 5, p. 1379, 2020.

[34] H. A. A. Al-Khamees, N. Al-A'araji, and E. S. Al-Shamery, "Classifying the human activities of sensor data using deep neural network," in *International Conference on Intelligent Systems and Pattern Recognition*. Springer, 2022, pp. 107–118.

[35] H. A. A. Al-Khamees, N. Al-A'Araji, and E. S. Al-Shamery, "An evolving fuzzy model to determine an optimal number of data stream clusters," *International Journal of Fuzzy Logic and Intelligent Systems*, vol. 22, no. 3, pp. 267–275, 2022.

[36] S. Vieira, W. H. L. Pinaya, R. Garcia-Dias, and A. Mechelli, "Deep neural networks," in *Machine Learning*. Elsevier, 2020, pp. 157–172.

[37] S. Dargan, M. Kumar, M. R. Ayyagari, and G. Kumar, "A survey of deep learning and its applications: a new paradigm to machine learning," *Archives of Computational Methods in Engineering*, vol. 27, no. 4, pp. 1071–1092, 2020.

[38] S. A. Alasadi and W. S. Bhaya, "Review of data preprocessing techniques in data mining," *Journal of Engineering and Applied Sciences*, vol. 12, no. 16, pp. 4102–4107, 2017.

[39] A. Rácz, D. Bajusz, and K. Héberger, "Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification," *Molecules*, vol. 26, no. 4, p. 1111, 2021.

[40] H. Mehmood, P. Kostakos, M. Cortes, T. Anagnostopoulos, S. Pirttikangas, and E. Gilman, "Concept drift adaptation techniques in distributed environment for real-world data streams," *Smart Cities*, vol. 4, no. 1, pp. 349–371, 2021.

[41] C. C. Aggarwal, *Data streams: models and algorithms*. Springer, 2007, vol. 31.

[42] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, and G. Zhang, "Learning under concept drift: A review," *IEEE Transactions on Knowledge and Data Engineering*, vol. 31, no. 12, pp. 2346–2363, 2018.

[43] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the kdd cup 99 data set," in *2009 IEEE symposium on computational intelligence for security and defense applications*. Ieee, 2009, pp. 1–6.

[44] R. Cattral, F. Oppacher, and D. Deugo, "Evolutionary data mining with automatic rule generalization," *Recent Advances in Computers, Computing and Communications*, vol. 1, no. 1, pp. 296–300, 2002.

[45] I. Katakis, G. Tsoumakas, E. Banos, N. Bassiliades, and I. Vlahavas, "An adaptive personalized news dissemination system," *Journal of intelligent information systems*, vol. 32, no. 2, pp. 191–212, 2009.

[46] J. Shao, Y. Tan, L. Gao, Q. Yang, C. Plant, and I. Assent, "Synchronization-based clustering on evolving data stream," *Information Sciences*, vol. 501, pp. 573–587, 2019.

[47] G. Ditzler, "Incremental learning of concept drift from imbalanced data," 2011.

[48] J. A. Blackard and D. J. Dean, "Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables," *Computers and electronics in agriculture*, vol. 24, no. 3, pp. 131–151, 1999.

[49] A. Vergara, S. Vembu, T. Ayhan, M. A. Ryan, M. L. Homer, and R. Huerta, "Chemical gas sensor drift compensation using classifier ensembles," *Sensors and Actuators B: Chemical*, vol. 166, pp. 320–329, 2012.

[50] V. Losing, B. Hammer, and H. Wersing, "Interactive online learning for obstacle classification on a mobile robot," in *2015 international joint conference on neural networks (ijcnn)*. IEEE, 2015, pp. 1–8.

[51] R. Chereshnev and A. Kertész-Farkas, "Hugadb: Human gait database for activity recognition from wearable inertial sensor networks," in *International Conference on Analysis of Images, Social Networks and Texts*. Springer, 2017, pp. 131–141.

Hussein A. A. Al-Khamees He received B.Sc. degree in computer science from the University of Babylon - Iraq in 1999, the M.S. degree in information technology from University of Turkish aeronautical association - institute of science and technology, Ankara, Turkey in 2017. In 2022, he received his Ph.D. degree in computer science from the University of Babylon. His main research interests are data mining, data stream analysis, deep learning and intelligent systems.

Nabeel Al-A'araji He received B.Sc. degree in mathematics from Al-Mustansiryah University, Iraq, in 1976, M.S. degrees in mathematics from University of Baghdad, Iraq, in 1978, and the Ph.D. degree in mathematics, from University of Wales, Aberystwyth, UK, in 1988. He is currently a professor in the software department, University of Babylon. His research interests include artificial intelligence, GIS machine learning, neural networks, deep learning and data mining.

Eman S. Al-Shamery She received the B.Sc. and M.S. degrees in Computer Science from the University of Babylon, Iraq, in 1998 and 2001, respectively. After completing her M.S., she worked as an assistant lecturer at the department of computer science, University of Babylon. In 2013, she received her Ph.D. in computer science from the University of Babylon. Currently, she is a professor in the software department, University of Babylon. Her current research interests include artificial intelligence, bioinformatics, machine learning, neural networks, deep learning and data mining.