An Adaptive and Late Multimodal Fusion Framework in Contextual Representation based Evidential Deep Learning Dempster-Shafer Theory

Doaa Mohey Eldin (✉ d.mohey@alumni.fci-cu.edu.eg)
faculty of computers and artificial intelligence
Aboul Ella Hassanein
faculty of computers and artificial intelligence
Ehab E Hassanien
faculty of computers and artificial intelligence

Research Article

Keywords: multimodal data fusion, modality, context-ware, late fusion, deep learning, uncertainty

Posted Date: November 28th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3653613/v1

Additional Declarations: No competing interests reported.

An Adaptive and Late Multimodal Fusion Framework in Contextual Representation based Evidential Deep Learning Dempster-Shafer Theory

Doaa Mohey El-Din[1,*,] , Aboul Ella Hassanein[2,3],*[b], Ehab E Hassanien[1]

[1]Faculty of Computers and Artificial intelligence, Cairo University, Cairo, Egypt

d.mohey@alumni.fci-cu.edu.eg (D. Mohey Eldin), aboitcairo@gmail.com (A. Hassanein), E.ezat@fci-cu.edu.eg .(E. Hassanien)[c]

Abstract: There is a growing interest in multidisciplinary research in multimodal synthesis technology to stimulate diversity of modal interpretation in different application contexts. The current literature review focuses on modality-based systems in a specific known context and leaves a gap in fusing multiple modality data types in different contexts. Therefore, there seems to be a real requirement for an analytical review of recent developments in the field of data fusion. The real requirement for modality diversity across multiple contextual representation fields is due to the conflicting nature of data in multi-target sensors, which introduces other obstacles including ambiguous, uncertain data, imbalance and redundancy in multi object classification. Additionally, there is lack of frameworks that can analyze offline stream data to identify hidden relationships between different modal data types and different modal counts. Additionally, the lack of a multimodal fusion model capable of determining the extraction conditions of the extracted fusion data has led to low accuracy rates in classifying objects across modalities and systems.

This paper proposes a new adaptive and late multimodal fusion framework to interpret multiple modalities and contextual representations using evidence-enhanced deep learning-based Dempster-Shafer theory. The proposed multimodal fusion framework is a MultiFusion learning model solution to solve the Modality-and context-based fusion to improve remote management, intelligent systems, and decision making. The proposed multimodal fusion framework can address the contradictory nature of data uncertainty, diversity of methods, factors, conditions, and relationships for multimodal explanation in multi-context systems to improve decision making and control in diverse contextual representations. Furthermore, this research provides a comparative analysis of the current fusion and prior multimodal data fusion models, explaining the differences of the construction analysis, mathematical analysis of fusion models, pros, and cons of them. In addition, this research presents a comparative analysis between the proposed framework with previous published fusion frameworks, exploring their concepts, advantages and problems, drivers, and current techniques. The experimental accuracy results in multimodalities experiments and multi-context using the proposed multimodal fusion framework is 98.45%. Additionally, some future research directions are discussed.

Keywords, multimodal data fusion, modality, context-ware, late fusion, deep learning, uncertainty.

1. Introduction

In recent years, the rapid advancements in artificial intelligence (AI) have paved the way for the application of deep learning techniques in various fields [1]. One prominent significant research field where deep learning has gained significant attention is decision-making processes [2]. One of depth dimension of researches is diving into the multi-fusion level framework utilized in decision-level using deep learning, aiming to provide a comprehensive analysis of its concepts, applications, and challenges.

Multi-sensor data fusion is a technology to enable inferring the information via multiple sources in order to form a unified full picture [3]. Multimodal data fusion (MMDF) is the process of combining disparate data streams (of different dimensionality, resolution, type, etc.) to generate information in a form that is more understandable or usable [4]. Multi-Data Fusion is the process of combining disparate data streams to generate information in a form that is more understandable or usable [5]. Data fusion is the joint analysis of multiple inter-related datasets that provide complementary views of the same phenomenon [6]. Data fusion systems are now widely used in various areas such as sensor networks [7], robotics [8], video and image processing [9], and intelligent system design [10], etc. The statistics of recent digital information around the world estimates 80%-90% of data generated by digitized services via industry is unstructured [10]. So, the data fusion becomes having a wide-ranging subject and important many terminologies have been used interchangeably.

These terminologies and ad hoc methods in a variety of scientific, engineering, management, and many other publications, shows the fact that the same concept has been studied repeatedly. The focus of this paper is on multi-sensor data fusion. Thus, throughout this paper the terms data fusion and multi-sensor data fusion are used interchangeably. The statistics between Number of Internet of Things (IoT) connected devices worldwide from 2019 to 2023, with forecasts from 2022 to 2030 that shows 51.11 billion sensors in 2025 and that expected to achieve 75.44 billion sensors in 2025. The data fusion research community have achieved substantial advances, especially in recent years [11]. Nevertheless, realizing a perfect emulation of the data fusion capacity of the human brainis still far from accomplished.

There is a revolution of using deep learning in multidata fusion [12]. First, single deep learning technique is used for interpreting single modality. Then, using single deep learning technique for interpreting the converted multiple modality topologies into one modality topology that causes missing some important data and not accurate enough. Either single modal alone is sufficient to predict the target accurately. That can interpret in the question "Can multi-modal learning provably perform better than unimodal?". Recently, research goes forward to use multi-deep learning techniques in different multi-topologies in diverse layouts based to save important data with higher accuracy results. That causes a high complexity specially most of research in one trained context. Recently, one of the open challenges in data fusion-based research is "*How to interpret multimodality in diverse topologies into multi-context without human intervention with high accuracy*?".

The main challenge of this paper is shown in "*Modality and context-based fusion*" in interpreting diverse data fusion for classifying objects and improving decisions from multi-targets into one unification objective for each system. The research challenge is entitled "Modality-Context-based Fusion". The research challenge is no work done on the combination between multiple modality-based fusions and multiple context-based fusions. All current literature review is based on context-based fusion or modality on specific known context [13], [14].

Furthermore, there is no fusion framework that can analyze offline stream data to generate concluded hidden relationship between different modality types of data and diverse modality numbers. Fusion problem is considered one of the most researched aspects of multimodal learning [15]. Fusion problem is also known by the science of heterogenous of interconnected data [16]. The open research of modality-context based fusion is shown in many problems properties as the followings, (1) *Standardization: a* hardness of the generalized context aware middleware due to the variety of contexts and systems involved to build a generic domain-focused middleware solution ~~[17]~~ [19]. (2) *Increase autonomy*: Although context aware middleware architectures minimize the requirement of human intervention when they serve personalized applications, human intervention is still necessary and playing a significant role in realizing context awareness [17]. (3) *Lack of testing:* It can be noted that most of middleware architectures contained in this paper are still at the conceptual stage [17]. (4) *Lack of accurate data*: due to different sources of problems, many times a context-aware system cannot build a computational model that represents the knowledge of a real-world domain [18]. Different fusion applications aimed to support of context representation and fusion, when formally incorporated in a context-aware system is still open research [19]. Approaches fuse the multi-modal features in a single way, which is not enough to elicit complementary data and then limits the performance [20].

The Modality-Context-based Fusion is considered a need to interpret modality input data types of specific fusion of diverse sensors that enable the discovery of various perspectives and objectives of each input of various perspectives and objectives of each input which are related to specific context models. The modality-context-based fusion causes conflicting of data nature to solve uncertainty, ambiguity, and Imbalanced interrelated data [21]. The open research challenges in fusion are shown as the following:

A) from various input sources [22, 23]. It is hard to build models that explores supplementary and not only complementary information and each modality might exhibit different types and different levels of noise at different points in time [24]. Most of the current AI research work focuses on disease discovery, classification and prediction from single modality data [25]. Fusion is the most common challenge in MML. Extracted information regarding diverse fusion types. The identification of the connection between modalities, concept-wise modality representation and tackling the ambiguity in high-dimensional data [26]. Multimodality is an interdisciplinary approach that considers communication and presentation as more than just language [27]. Modality is an important concept in intelligent systems, referring to the process of combining information from different input methods. The goal is to improve the accuracy of classification and prediction results by integrating several methods. This integration is important in several recent smart systems/applications, e.g., smart home, smart health, smart agriculture, and smart mobility. The method-based fusion problem poses an important challenge in smart environment applications, and it is essential to solve the fusion problems in context- and modality-aware smart environments. This problem is determined by the interpretation of the input data, which can be classified into two types: same type of data, such as image only, text only, etc. and different types of data, such as image-text, text-audio, image-video, etc.

B) Contextual data can come from a variety of sources such as sensors, wearables, and social media, which can introduce noise and inconsistencies in the data [28]. Context awareness is an important feature of recent intelligent systems that allows the system to adapt to changing environments and provide personalized services to users [29]. It refers to the ability of a system to detect, interpret, and respond to contextual information related to users, devices, locations, and events. Context awareness has been widely used in various intelligent systems/applications. The problem of context-based unification arises from the difficulty in interpreting the different contexts that exist in intelligent systems/applications. It is important to develop effective data processing and aggregation techniques to ensure contextual data is accurate and reliable. Scalability and adaptability are related to the key issue of the combination of modality and context of context-aware systems. The lack of formality and generality in the previous context representation models [30]. Most of the current artificial intelligence research work concentrates on disease discovery, classification and prediction from single modality data [31]. It was hard to infer relationships among different information and infer meaningful results from it. The main goal of all this process was to come up with the complete picture and understand how different factors were connected and affected each other. The term context awareness describes the ability of a system to integrate data based on its specific context, taking into account the timing of data flows and the dynamic nature of context-specific data samples. Implementing context awareness in smart environments faces a number of challenges. One of the main challenges concerns the accuracy and reliability of contextual data collection.

C) Conflicting Data

Moreover, there is no way to solve the conflicting nature of data (Image, Text, Audio, and Video) in multi-target sensors to object classification in diverse context systems, heterogenous data, imbalanced data, unstructured data, conflicting data, different representation, difficult modality numbers [32]. The research challenge of "Contextual Method-Based Fusion" stems from many issues [33]. As follows, the difficulty in finding commonalities between smart systems requires common elements in the connectivity of smart devices. Mismatch in input method type, input target, and input relationship between different intelligent systems. Lack of a single intelligent system dataset capable of understanding a large number of inputs for testing, leads to the need to test multiple intelligent systems with different inputs and different numbers of datasets. The field of research is called "classification-based fusion" and has illustrated the difficulty of interpreting object classification across a variety of data types in a variety of intelligent environments in offline streams.

This research paper presents a new adaptive and late multimodal fusion framework that relies on creating a MultiFusion learning model for solving modality-context-based fusion challenges for improving decision-making and controlling systems. The proposed adaptive fusion framework creates fully automated in selective deep neural network and constructing adaptive fusion model for all modalities types based on input type. The proposed framework is constructing automatically deep neural network of the Dempster-Shafer and concatenation to achieve to a bigger number of features for interpreting unstructured multimodality types based on late fusion. The proposed framework is implemented based on five layers, a Software-Defined fusion layer, a preprocessing layer, a dynamic classification layer, an Adaptive fusion layer, and an Evaluation layer. It is formalized the modality-context-based problem into an adaptive multimodal mathematical fusion framework based on late fusion level.

The rest of this paper is organized as follows: in Section 2 the background of data fusion, definitions, levels, techniques, conceptualizations, methodologies, and purposes, as well as the major benefits of data fusion, and limitations are explored. The related works of modality fusion problem and context-aware fusion problem performing data fusion are discussed in Section 3. Section 4 provides a proposed solution of modality-context based fusion that presents a new Adaptive Multimodal Fusion framework based on data perspective Fusion Taxonomy. A proposed multimodal fusion framework presents a solution for reusable to support the Multi-Fusion Models using Deep Learning and Dempster-Shafer Theory for same and different modality types. This section is explored a comparative analysis of the proposed multimodal framework and previous frameworks. In Section 5, experiments and results with respect to multimodality and Multicontext, are provided. It presents a comparative analysis between proposed fusion model and previous Fusion models. Finally, Section 6 presents the concluding outlines for this paper and future works.

2. Literature review

2.1.1 Data Fusion and Multimodal data fusion

Data Fusion is the process of combining information from heterogeneous sources into a single composite picture of the relevant process, such that the composite picture is generally more accurate and complete than that derived from any single source alone [33]. Multi-Data Fusion: is the process of combining disparate data streams to generate information in a form that is more understandable or usable [34]. Multi-Data Fusion consists of the combining between two definitions of ult-sensor data fusion and multimodal data fusion. Multisensor data fusion (MSDF): is the process of combining observations from a number of different sensors to provide a robust and complete description of an environment or process of interest. Multimodal data fusion (MMDF) is the process of combining disparate data streams (of different dimensionality, resolution, type, etc.) to generate information in a form that is more understandable or usable [35].

2.1.2Data Fusion Approaches

The data fusion approaches have three types as shown in Figure.2, early fusion, late fusion, and hybrid fusion as shown the differences between them in table.1. In the first type of fusion a raw data from different methods is combined at the input level before being fed into the model, [36]. While is the Late fusion for the data from each method is processed independently through separate models, and the results of these models are then combined at a later stage [37], [38]. Third type of fusion that combines different fusion strategies to achieve the desired results [39], [40].

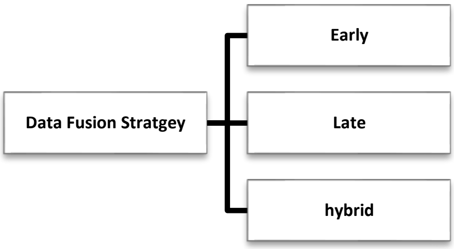


Figure.2: Data Fusion Strategies Levels

Table.1: A comparative analysis of fusion levels.

No.	Fusion Level	Methodology	Advantages	Disadvantages
1.	**Decision/Late fusion**	Fusion of multiple inputs after classifications	It enhances the recognition of complex real-world activities and derives the detection probability for improving accuracy and performance.	It applies on depression of the big parameters of modalities.
2.	**Feature/ Early Fusion**	Fusing multiple inputs with all features.	fusing big features from various models	not reach suitable accuracy enough
3.	**Hybrid Fusion**	Using fusion techniques for Interpreting and fusing multiple inputs with all redundant features and parameters.	It improves accuracy and performance	The presented solution is still low efficiency. And is high redundant fused data and high complexity

2.3 Data Fusion Techniques

Data fusion techniques can fuse extracted data via multiple intelligent devices or sensors [3] and relative metadata from databases to reach enhanced accuracy results. The important fusion techniques are Central Limit Theorem (CLT), Kalman-Filter (KF), Bayesian Networks (BN), Dempster-Shafer Theory (DST), and Deep Learning (DL) Algorithms**,** as shown in the following. Table 2.2 shows a comparative study between data fusion Techniques based on SWOT Analysis. The importance of deep learning, transfer learning, and reinforcement deep learning techniques are shown in the classification to improve the fusion results as used in the decision fusion level.

2.3.1 Central Limit Theorem (CLT)

CLT refers to the understanding several variables [30]. For graphing n that describes the population's random variables according to mean µy and σ. y is explaining for the standard deviation. The mean distribution can equal the mean (µŷ) for the Population Mean (µy) and expresses the standard deviation (σŷ) for the population standard deviation. It is divided by taking into account the square root (σy) of a sample number, n. Z score for the sample. X is the standardized mean score, and σ presents the standard deviation as applied on pervious equation (1) and equation (2).

$$Z = \frac{(x-\mu)}{\sigma}, \qquad (1)$$

$$Z = \frac{(x-\mu)}{\sigma/n}, \qquad (2)$$

2.3.2 A Kalman-Filter (KF)

Kalman filter is an estimation algorithm for explaining the status of a separated time-controlled operation depicted by a linear stochastic equation. KF fuses all information [32]. Its procedures the ready measurements equation and ignores the precision to evaluate the current value of the advantage variables. It supports estimating the current features utilizing current measurements and prior parameter measures as shown in equations (3 and 4).

$$X(K) = Ax_{k-1} + B_k u_k + w_{k-1}, \qquad (3)$$

$$Z_K = \ H_k x_k + y_k\,, \qquad (4)$$

Where $x_k$is System state vector, W_k is the process noise, y_k is the measurement nice. H is the estimated transfer matrix, as is a transition matrix, and B relates to manage the input.

2.3.3 Bayesian Networks

Bayesian networks produce data fusion measurement; they are a mutual method applied for multi-sensor data fusion in the static environment [42]. The probability distributions can examine a convenient to suspicious data processing based on the added noise of Gaussian. If any noise can influence a multi-sensor data fusion system, this doesn't apply to deduce and save the original data. Kalman-filter (KF) depends on a pure mathematics approach for the problem-solving and getting analytics. The major idea of data fusion is the fusion of data established on the uncertainties. It aims to identify targets and parameters judgment. According to the entire procedure, and the probability function plays a significant role. The probability function looks like the reverse of conditional probability. The parameter is represented as θ, the event probability A denoted as applied in pervious equations (5 and 6).

$$P(A|\Theta) = \frac{P\,(A,\Theta)}{P(\Theta)}, \qquad (5)$$

$$P(A|\Theta) = \frac{P\,(A|\Theta)P(\Theta)}{P(A)}, \qquad (6)$$

According to equations 5, 6, it is significant to identify the values of θ if it is unknown parameter which requires knowing several metadata with some measurements. Assuming θ can examine the total distribution parameter (A) θ χ ρ.

2.3.4 Dempster-Shafer Theory (DST):

DST relies on the Bayesian theory that explains the canonical approach for statistical inference challenges. The Dempster-Shafer decision theory becomes a generalized Bayesian theory [7], [34]. It eases the evaluation of proposition distribution and the union propositions. The Dempster-Shafer is very powerful in the causes system recognizing the total mutual context facts of the same type in "the frame of discernment θ" as mention in pervious equation (7).

$$\theta = \{A, B, \{A, B\}, \{somebody\ else\}\}, \qquad (7)$$

The interpretation meaning of this example this person is "user-A", "user-B", "either user-A or user-B", or "neither user-A nor user-B, must be somebody else" 1. Each sensor, sensor *Si*, for instance, will participate in its notice by specifying its beliefs over Θ. The function is known as the "probability mass function" of the sensor *Si,* indicated by mi. So, with respect to sensor Si's notice, the probability which "the detected person is user A" is specified by a "confident interval," as illustrated in equation (8).

[belief, (A), $\boldsymbol{Plausibility}$ $_i$ (A)], *(8)*

The lower bound of the confidence interval which refers to the belief confidence (as explained in equation (9), that has for all evidence E_k that can help the given proposition "user A". The plausibility confidence is considered the upper level of the confidence interval, and it can compute them by the given proposition.

$$(m_i \oplus m_j)(A) = \frac{\sum_{E_i \cap E_j = \emptyset} m_i(E_k) m_j(E_k)}{\sum_{E_i \cap E_j = \emptyset} m_i(E_k) m_j(E_k)}, \qquad (9)$$

2.3.5 Deep Learning Algorithms: The counted number of needed interconnections for example the constructed network generates unreasonably fast as the size of the input developments [20], [27]. The comparison of fusion algorithms is shown in sensor fusion based on the SWOT Analysis. There are many techniques of deep learning such as CNN, RNN, ANN, and Transfer learning,

2.4 A Comparative Analysis between Proposed Adaptive fusion model and Pervious Fusion models

This section will discuss a comparative analysis between the proposed adaptive fusion model and two prior fusion models [41] and [42] as shown in Table.2. This comparison relies on the of the differences between multimodal fusion models' properties of the modality data type, modality number, data fusion level, interpreted context considerable, experimental dataset, and weaknesses.

Table.2: A comparative analysis between proposed Adaptive Fusion Model& Previous Fusion Model.

Research paper	Multimodal Interfaces	Development	Input modality type	Context	Technique	Adaptive to any Context	Advantages	Weakness
(Vaezi et al., 2020) [41]	Early	Feature	Audio-visual	Human Action Recognition	Hierarchical feature fusion	hardness to multiple models' types	86%	hardness to multiple models' types
(Tong et al., 2021) [42]	Late	Decision	Image dataset	CIFAR-10	Automated fusion	Can't apply on multiple context	More than 89%-94%	Can't apply on multiple context
Proposed adaptive multimodal fusion model of proposed framework, 2023	Late	Decision	Multiple modalities	Multiple context	Deep learning improved Dempster-Shafer fusion	Adaptive to multiple contexts (Classification objects)	95% - 98%	Complex

2.5 A Comparative Analysis between Proposed Adaptive Multimodal Framework and Pervious Multimodal Fusion Frameworks

This section will discuss a comparative analysis between the proposed adaptive framework between three multimodal frameworks [43], [44], and [45] as shown in Table.3. This comparison relies on the of the differences between multimodal framework's properties of the modality data type & modality number, data fusion level, interpreted context considerable, experimental dataset, and weaknesses.

The proposed adaptive framework can solve many previous drawbacks in [43], [44], and [45]. The three previous frameworks can't interpret multimodality input in diverse contexts for improve object classification. Its advantages are shown in the interpretation of multimodality types and multimodality number dynamically (based on data perspective not context perspective). It can excavate the relationship between multimodalities. It can control automatically the same multimodality and various multimodality of (Text, audio, image, and video). It can reach high accuracy of classification of one / multi object classification. In addition, the proposed adaptive framework can solve the redundancy fused data problem (redundant data) and high level of abstract data problem that is based on low features. It can remove redundancy of fused vectors data. It is designed based on deep neural network models with fusion of the Dempster-Shafer fusion with Concatenation fusion. It has development implementation and user interface with high complex of implementation.

Researchers in 2017 [43] presented Adaptive Multimodal Environment (FAME) that can make a fusion of Mobile sensors based on (gesture, speech, GPS) with respect to location-based dimension. It can't be adaptive to multiple contexts but it has one consideration of One context location-based Augmented reality system. It didn't have a development implementation and didn't have multimodal interface.

Researchers in 2022[44], presented Adaptive Multimodal Emotion Detection Framework that can deal with Three input modalities only, images, text, audio, on one context of the Facial robots. It is designed based on early fusion level that is constructed by the deep neural network and concatenation. Although it can improve Tri-modality interpretation in one context, it is still facing a problem in multimodality interpretation with multiple contexts. It is developed but it does not have a multimodal interface. Researchers in 2015 [45], presented ModDrop: adaptive multi-modal gesture recognition that relies on the bimodality interpretation (Video and speech) of one context of the

Table.3: A comparative analysis between proposed Adaptive Fusion framework & Previous Fusion Frameworks

Multimodal fusion framework	**Development**	**Input modality type**	**Context**	**Technique**	**Adaptive to any Context**	**Advantages**	**Weakness**	**Multimodal Interfaces**	**Development**
Adaptive Multimodal Environment (FAME) [43], 2017	No	No	Mobile sensors based on (gesture, speech, GPS)	One context location-based Augmented reality system	Fusion of dialog engine	Fusion and fission	No.	Try to improve accuracy in one context to multimodality with respect location	Not implemented Not suitable to multiple contexts
Adaptive Multimodal Emotion Detection Framework [44], 2022	No	Yes	Three modalities, images, text, audio, on Facial robots	One context Hand writing dataset and audio dataset	Deep neural network and concatenation	Early fusion	No	Improve Tri-modality interpretation in one context	Not suitable to multimodality types in same and different types Not suitable to multiple contexts
ModDrop: adaptive multi-modal gesture recognition [45], 2015	No	Yes	gesture of human reaction based on spatial scale (Video & Speech)	One context	Deep neural network (multi-scale)	Late fusion	No.	Fused bimodality in one context with multiscale learning model in one context.	Not suitable to multiple modalities and not applicable on multiple contexts
Proposed adaptive multimodal framework, 2023	Yes	Yes	Text, audio, image, video, (Different and same number)	Any context	Deep neural network models with adaptive fusion of Dempster-Shafer fusion with Concatenation fusion	Late	Yes. Be applicable in diverse smart systems such as, Smart health, & smart military.	-Interpret modality types and modality number dynamically (based on data perspective not context perspective). Excavate the relationship between modalities. -Reach high accuracy of classification of one / multi object classification. -Remove redundancy of fused data	High complex implementation

gesture of human reaction based on spatial scale. It is designed based on late fusion level. It is powerful bimodality in multiscale. It is not suitable to multiple modalities and not applicable on multiple contexts.

3. The proposed Adaptive Multimodal Fusion Framework

3.1 Adaptive multimodal fusion framework Architecture

The adaptive multimodal fusion framework is designed based on a fully automated controlled of the combination of deep neural networks and two fusion levels for improving object classification accuracy. Model-based fusion is fusion level one and Feature-based fusion is second fusion level. The proposed framework is constructing automatically deep neural network of the Dempster-Shafer and concatenation to achieve to a bigger number of features for interpreting unstructured multimodality types based on late fusion. The proposed framework is implemented based on five layers, a Software-Defined fusion layer, a preprocessing layer, a dynamic classification layer, an Adaptive fusion layer, and an Evaluation layer as shown in Figure.1.

The proposed framework formalizes the modality-context-based problem into an adaptive multimodal mathematical fusion framework.

- (How) How interpret the multimodal data?
- (When) What are conditions multimodal performs better than unimodal?
- (Where) How can achieve the fusion regardless the context?
- (How) How can make fully automated multimodality fusion in multiple contexts?
- (Why) What results in the accuracy gains?
- (How) How result is Unique?

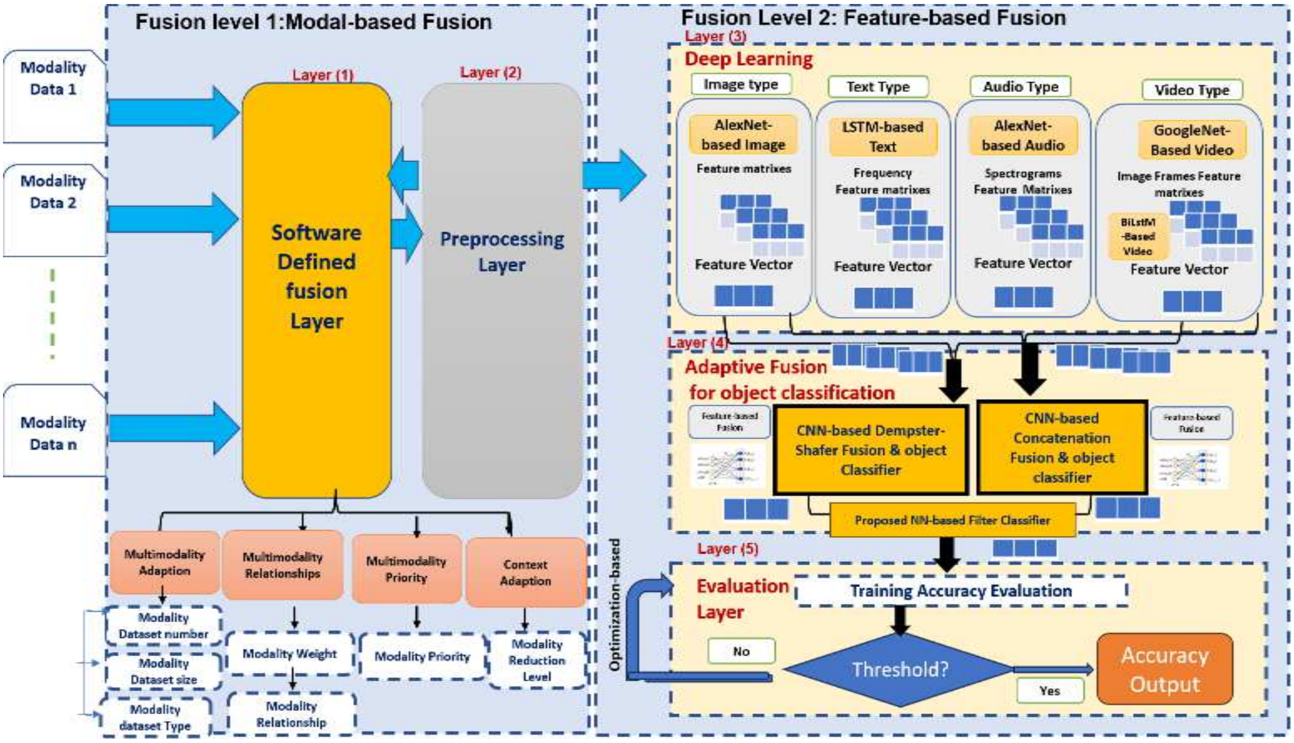


Figure.1: An Adaptive Multimodal Fusion Framework in Contextual Representation based on Late Fusion Level using MultiFusion Learning Model and Improved Evidential Deep Learning Dempster-Shafer

The proposed solution is building up a versatile multimodal fusion system that depends on making a learning show for tackling modality-context-based combination challenges for moving forward decision-making and controlling frameworks. The proposed versatile multimodal combination system makes completely computerized in particular profound neural arrange and developing versatile fusion demonstrate for all modalities sorts based on input sort. It is designed based on late fusion level. A Proposed adaptive multimodal fusion framework plays important role in realizing multi object classification in management smart systems. The goal of study Smart environment is not considerable specific context but it refers to interpreting the common big data characteristics of extracted data via multiple smart devices. The proposed solution has achievement of the fusion Objective that can fuse the interrelated of the complementary data for mange the data with unification objective.

Input: multimodalities datasets number.

Proposed solution: proposed deep learning with improved dempster-Shafer fusion.

Output: classified fused vector to target Y of the objects in interrelated data. The output is Unique and Optimized-based for achieving the best point. It achieves to high accuracy in multiple modalities and multiple contexts.

Tracing: based on data modality input (the tracing may be an image or an interval of features that can convert into numerical vectors and check classifier with proposed output vector of all features of all objects).

The proposed framework architecture is designed based on two fusion levels as shown in Table.4.

Table.4: A proposed multimodal fusion framework based on two fusion levels (Inputs/output)

Input	**Modality dataset input (same/different), such as (Image, Text, Audio, Video)**			
Fusion level 1: Model-based Fusion				
	Input	Processing		Output
Layer 1	Modalities dataset	Control multimodality input number and type	Interpret weights, priorities, and dataset size.	Model/process-based fusion level of the multimodality to deep learning
Layer 2	Input	Processing		
	Interpret modality type, number, computed weights, and computed priorities.	Preprocessing based on multimodality input		
Fusion Level 2: Feature-based Fusion				
	Input	Processing	Output	
Layer 3	After Processed modality datasets (and four parameters: Interpret modality type, number, computed weights, and computed priorities).	Deep learning based on modality types	Reduced feature matrixes from various modalities Convert feature learned matrixes into vectors Convert feature non-learned matrixes into vectors	
Layer 4	Feature learned vectors for each data type from deep learning based on modality number (for dempster-Shafer) Feature non-learned vectors each data type from deep learning based on modality number (for concatenation)	Adaptive fusion model based on deep learning of automated dempster-Shafer and concatenation & Feature vector of priority and weight	Automated object classifier feature fusion of dempster-Shafer based on computed belief and plausibility to show the one fused feature vector for all objects in modality dataset.	Automated object classifier feature fusion of concatenation fusion based on fused all matrices in one fused feature vectors for all objects in modality datasets.
	Proposed NN-based Classifier for fused vectors from two fusion models.			
	Make one fused filtered vector from two fused vectors with remove redundancy vectors.			
Layer 5	Input	Processing	Output	
	Fused feature Vector	Measure accuracy of classified fused feature vector	Accuracy condition	
		If the accuracy is satisfied threshold?		
		Yes	Fused Feature Multi object vectors	Accuracy matrix
		No	Training on changing hyper parameter to achieve best fit accuracy point	Again, tracing of deep learning in layer (3) to achieve satisfied accuracy
Output	**Fused Feature Multi object Vector**		**Accuracy confusion matrix**	

3.2 Preliminaries

A) Preparation of proposed dataset criteria

The suitable datasets can be applied on the presented framework to achieve the generic and adaptivity of framework that have one criteria of mutual properties as shown in Table.5.

Table.5: A proposed one data criteria based on mutual propoerties

Proposed dataset criteria		
Heterogeneous Data characteristics whether same or different data	Unstructured Data	Different representation
Balanced or Imbalanced Data (different dataset size)	Different modality dataset size	Interrelated data based on type or metadata about something
Time series data	Different modality dataset number	Different characteristics
Diverse sources	Conflicting Data topologies	Big data interpreted
Supervised data	Unknown context	No moving window based on time.

Time: Not is effect on the output results but it is not based on the fusion (no moving window).

The suitable datasets can be applied on the proposed criteria has two types:

1. Fusing similar types of data from various sources to object classification. For example (thermal datasets of weapons to achieve the best full vision of each weapons object's).
2. Fusing multi-target data via diverse types or different characteristics to achieve the unification target of object classification that are often interrelated to each other. For example (fusing multiple patients records Excel sheet with X-rays to make profiling of patient and profiling diseases from the data).

B) preparation of proposed pre-conditions

The pre-conditions of the input data criteria for using the proposed framework to achieve the generic and adaptivity of framework targets as shown in Table.5. Preconditions allows functions to provide minimum required values for the modality input datasets:

1. Datasets aim to object classification.
2. Datasets are interrelated data.
3. Unknown context domain.
4. Maximum number of modality input number from 1 to 16 (limit of tracing).
5. The modality dataset images are based on various 2D images and not suitable to 3D images.
6. Sound is sound classifier: is not talk (classifier sound based on signal features of the frequency, high, speed, and time).
7. Video may be long or short stream (video has objects to classify).

3.3 The Proposed MultiFusion Learning Model Description

The proposed MultiFusion learning model is designed to solve modality-context-based fusion challenge as shown the description of each problem modality-based fusion and context-based fusion in Table.6. The proposed MultiFusion Learning model is designed based on two fusion levels, Model-based fusion and Feature-based fusion as shown in Table.6. The proposed MultiFusion learning model presents a proposed correlation between multiple datasets inputs proofed mathematical controlling for different topologies and a proposed classification model between multimodalities dataset into unification feature matrixes with reduction feature level of objects via modality inputs.

Table.6: A proposed solution achievemenet twotarget of problems.

Modality-context-based fusion challenge	**Solution Achievement**
Modality-based fusion	The modality-free defines multimodality with tracing on 1-16 training input of all modalities number with respect on interrelated objects.
Context-based fusion	The context-free defines the space of all responses allowed for the given computation and relationship between all parameters in different modalities input to interrelated objects. The modality defines the adaption of automated multi object classification. The proposed modality adaptation technique not only learns common knowledge between categories, but also learns additional relative knowledge from each category. The proposed modality adaptive approach also incorporates knowledge of the label space. The proposed approach is a complete architecture where all parameters are updated in one step.

Table.7: The description of two fusion levels based on proposed layers

	Name	**Included**	**Goal**
Fusion level one:	**Model based fusion**	Layer 1: Software-defined fusion layer Layer 2: preprocessing	Count modality data type Modality dataset size Modality data number of each type Extract new correlation between modalities dataset input based on weight, priority, reduction level, and extracted relationship Evaluate mathematical proofed weight and priority dynamically
	A proposed correlation between multiple datasets inputs proofed mathematical controlling for different topologies with extracted proposed correlation with multimodality in Multicontext (unknown context). The preprocessed data has a changed measurement of proposed correlation based on computed weight and priority		
Fusion level Two:	**Feature based fusion**	Layer 3: Deep learning layer Layer 4: adaptive fusion layer Layer 5: Accuracy evaluation.	Deep learning of different topologies into one topology of reduced matrixes. adaptive fusion with improved dempster-Shafer technique. Accuracy evaluation.
	A proposed classification learning model between modalities dataset input into unification matrixes topologies with reduced features of all objects in the datasets. It consists of three layers that are entitled, dynamic classification layer, adaptive fusion layer, and evaluation layer. It evaluates the accuracy optimization using swarm optimizer to achieve the best fit point of accuracy.		

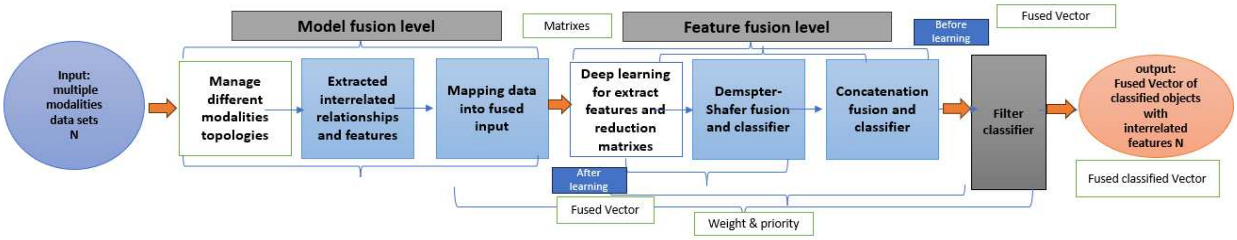

Model fusion level
Matrixes
Feature fusion level
Before learning
Fused Vector
Input: multiple modalities data sets N
Manage different modalities topologies
Extracted interrelated relationships and features
Mapping data into fused input
Deep learning for extract features and reduction matrixes
Demspter-Shafer fusion and classifier
Concatenation fusion and classifier
Filter classifier
output: Fused Vector of classified objects with interrelated features N
After learning
Fused Vector
Weight & priority
Fused classified Vector

Figure.2: Abstracted multimodal fusion framework (based on Iput/output).

A) *Fusion level one: Model/process-based fusion*

The first fusion level is designed based on model-based fusion. This level interprets multiple topologies with different modalities and diverse characteristics. This level extracts new correlation between modalities dataset input based on weight, priority, reduction level, and extracted relationship. It consists of two layers that are entitled, software defined fusion layer, and preprocessing layer. It evaluates the mathematical proofed weight and priority dynamically, count modality data type. It measures the modality dataset size, Modality data number of each type.

Layer (1): Software-Defined fusion layer: is considered as a controller for creating a proposed correlation between multiple datasets inputs. This layer can control the input based on modality types whether Unimodal, bimodal, tri-modal, or more multiple modals and modality number with diverse features whether same modality types or different modality types. A Software-Defined fusion layer is constructed based on five dimensions, modalities data type, modality data number, modality dataset size, the weight of features relationships interpretation, and relationship weights of priority of each modality. A Software-Defined fusion is extended from the Software-Defined terminology which refers to the software-controller or the management of application programming interface (API) such as Software-Defined network. A Software-Defined fusion layer is constructed based on the following issues,

A) Multimodality Adaption for multiple modality inputs: refers to the dynamic modality number of input, type, and size of interrelated data. Interpret the inference of four modality data (image, text, audio, and video). Dealing with multiple input numbers as defined by Equation (10) and its proof (1) as shown in Table.8.

$$\boldsymbol{I(n) = \sum_{i=1}^{N} \sum_{j=1}^{x} Dt_{xN}} \qquad \textit{\textbf{(10) Proof (1)}}$$

Table.8: Proof (1): (Interpretation Modalities Input.)

Let DT➔ data type, let n= the total number of modality inputs,
Let $\mathbf{Dt_x}$➔ a type of degree of each data, $\mathbf{Dt_N}$➔ total number of inputs of each type

Input= DT#, DTx

$\mathbf{I(n) = \sum_{i=1}^{N} \sum_{j=1}^{x} Dt_{xN}}$
$\mathbf{I(n) = \sum_{x}^{N} DT_{x1} + \sum_{x}^{N} DT_{x2} + \sum_{x}^{N} DT_{x3} + \sum_{x}^{N} DT_{x4}}$

Let $\mathbf{DT_{x1}}$➔ Image data type, $\mathbf{DT_{x2}}$➔ Audio Data type, $\mathbf{DT_{x3}}$➔ Text data type, $\mathbf{DT_{x4}}$➔ Video

$\mathbf{\sum_{i=1}^{N} \sum_{j=1}^{x} Dt_{xN} = \sum_{x}^{N} DT_{x1} + \sum_{x}^{N} DT_{x2} + \sum_{x}^{N} DT_{x3} + \sum_{x}^{N} DT_{x4}}$

The main question of proof (1) (Equation (1)) is: What is the number of modality datasets input and the type of each one? Due to the context is unknown and the dataset is dynamic, proof (1) aims to identify the number of modality dataset types and number of each type.

Example#1: the dataset has two images' datasets via two sources of (Thermal and RGB of object weapons) and each dataset has folder of number of images (Thermal Image Folder) and (RGB Folder). The DTx, x=1 1=image and DTn= 2.

Example#2: the dataset has three images' datasets via three sources of (Patient Metadata Excel sheets, patient additional Excel sheets and X-ryas of object medical) and each dataset has folder of number of images and one or excel sheets (Thermal Image Folder) and (RGB Folder). The DTx, x=3 , 3=Text and DTn= 2, and DTx, x=1, and DTn=1. So, $\sum_{i=1}^{N} \sum_{j=1}^{x} \mathrm{Dt}_{xN}$ = (n=3 number of datasets and type of two text and the third is Image).

B) Multimodality relationships Weight and Type: refers to the weight value of all modality data number and data size

It interprets the inference modalities in relationships. The weight factor of each dataset is computed based on the relationships between each dataset and the neighbor dataset as defined by Equation (11) and its proof (2) as shown in Table.9.

$$I(w)=\frac{\sum_{1}^{n} Dt_{x1N1}}{\sum_{x}^{N} Ds_{xN}} \qquad \textit{(11) Proof (2)}$$

Table.9: Proof (2): (Modalities Input Relationships)

```
Let DT➔ data type, let n= the total number of modality inputs,
Let $\mathbf{Ds_x}$➔ total size of neighbor biggest datasets,

Let $\mathbf{Ds_{xc}}$➔ Number of dataset size, $\mathbf{Dt_{Nc}}$➔ current dataset
                                        Relationships (1-1, 1-m, m-m)
Input= $\mathbf{Dt_{xcNc}}$, $\mathbf{Ds_{xN}}$
If ($\mathbf{Dt_{xcNc}}$, < $\mathbf{Ds_{xN}}$ )
{
I(w)= $\frac{\sum_1^n Dt_{xcNc}}{\sum_x^N Ds_{xN}}$
c(w)= $\left|\frac{1}{I(w)}\right|$ ,                let c(w)= computed weight of relationships (3)
}
Else if   ($\mathbf{Dt_{xcNc}}$, >= $\mathbf{Ds_{xN}}$ )
{
C(w)= I(w)
}
 End if
```

The default weight factor is computed for each modality input dataset size based on the division of the current dataset and the biggest dataset size. The main question of proof (2) (Equation (11)) is: How can compute the weight of each modality dataset and how can have an impact of the relationship of extracted features? Due to no previous information about the conditions of object classification, the main goal of extracted weight is counting the number of each dataset size and calculate the relationship of all dataset's sizes.

Example#3: the dataset has two images' datasets via two sources of (Thermal and RGB of object weapons) and each dataset has folder of number of images (Thermal Image Folder) and (RGB Folder). The DTx, x=1 1=image and DTn= 2. It has two cases: 1) If dataset size is equal

Dataset size (x=1)= 1000 images

Dataset size (x=2)= 1000 images.

$I(w)=\frac{\sum_1^n \mathbf{Dt_{xcNc}}}{\sum_x^N \mathbf{Ds_{xN}}}$ = $I(w)=\frac{1000}{1000}$=1 (each dataset is equal size to interpret in the interval in between 0 and 1that equals [1/number of datasets➔ each dataset has the same weight 1 and 1].With respect the interval of [0,1]

1) If dataset size is not the same size

Dataset size (x=1)= 1000 images

Dataset size (x=2)= 2000 images.

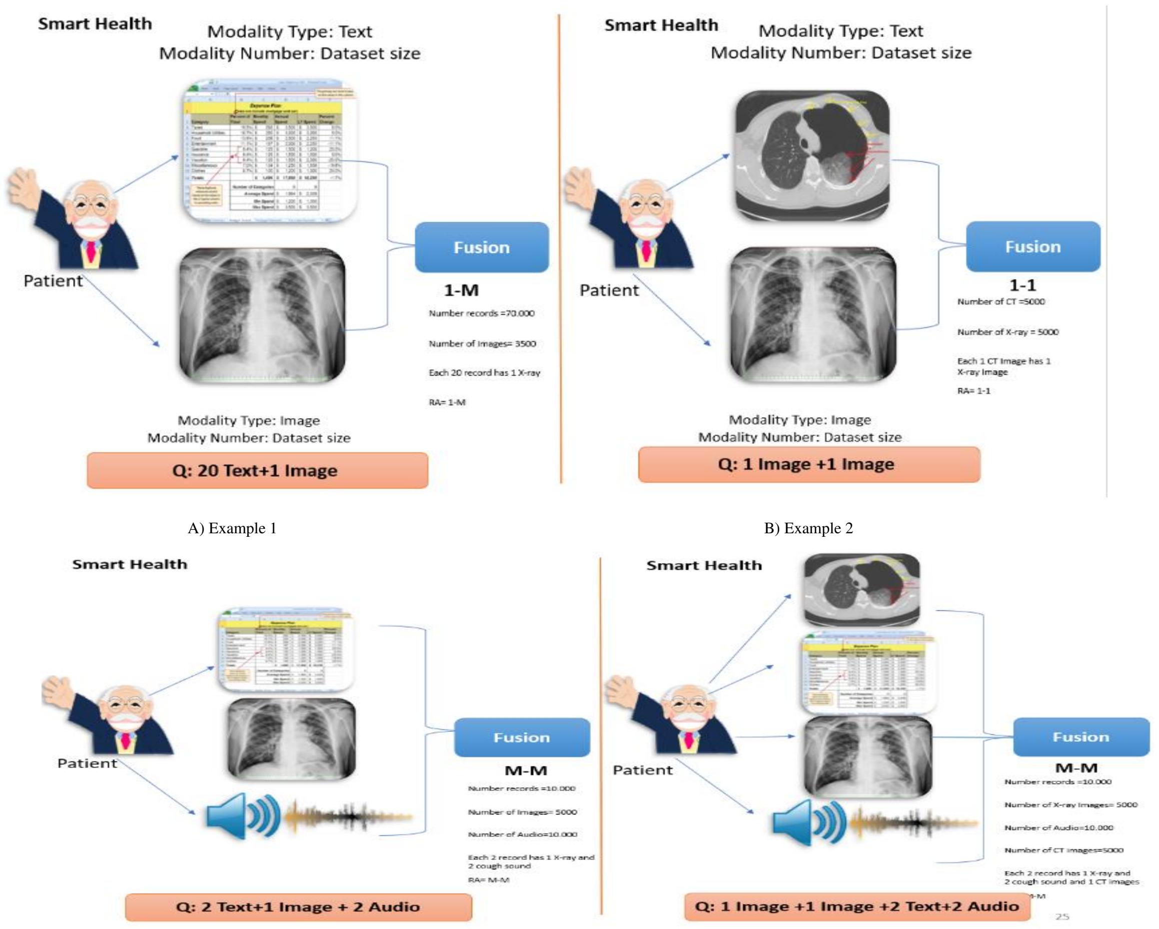
$I(w)=\frac{\sum_1^n \mathbf{Dt_{xcNc}}}{\sum_x^N \mathbf{Ds_{xN}}}$ = $I(w)=\frac{1000}{2000}$=1/2

(That interprets each two images are relative into one image of the second dataset, weight dataset1=1/2 and the second dataset size2= 1]).

Note: There has limit of interval relative data to fuse data from [1 to 1/10] of data. The relative limit can't fuse data less than related data. The importance of weighted value is shown more detailed of the interrelated data to improve the complementary data that is shown in the relevant and frequency of data.

Example#4, if dataset size of (x=1)=100.000 images to Dataset size (x=2)= 10.000 images. In other hand, what is important of 500 images classification with 100.000 images dataset and the hardness of fusing between details and accuracies and features.

Example#5: the dataset has three images' datasets via three sources of (Patient Metadata Excel sheets, patient additional Excel sheets and X-ryas of object medical) and each dataset has folder of number of images and one excel sheets (Thermal Image Folder) and (RGB Folder). The DTx, x=3 , 3=Text and DTn= 2, and DTx, x=1, and DTn=1. So, $\sum_{i=1}^{N}\sum_{j=1}^{x} \mathrm{Dt}_{xN}$ = (n=3 number of datasets and type of two text and the third is Image) that applies on the proofed Equation (2) interprets the relationship based on the size of dataset and the relevant total number of datasets. The modality relationship is shown in balanced or imbalanced data that can respect the relative number of dataset size that can interpret from three types One-to-One (1-1) relationship, One-to-Many (1-M) relationship, and Many-to-Many (M-M) relationship as shown in Figure.3 in the Example (1), Example (2), Example (3), and Example (4).



A) Example 1

B) Example 2

C) Example 3

D) Example 4

Figure.3: Fusion Examples for the same patient of classification disease in Smart Health based on various Inputs of Modality Types and Number.

C) Multimodality Priority: refers to the Importance of modality with respect to all datasets

It is based on a relationship between each data set with the lowest dataset. This priority relies on the subtraction of each modality dataset size and the lowest modality dataset size that divides by the summation of the total size number of all input modalities as proven in equation (12) as proven by Proof (3) as shown in Table.10.

$$p(f)=\frac{\sum_{1}^{n} Dt_{xcNc-} \sum_{1}^{n} Dt_{xlNL}}{\sum_{x}^{N} Ds_{SN}} \qquad \textbf{\textit{(12), Proof (3)}}$$

Table.10: Proof (3): (Priority of Each Modality Input)

Let DT➔ data type, let N= the total number of modality inputs,

S[]= {s1,s2,…….,sn} , S➔ the sizes of all input modality datasets

Let $\mathbf{DT_c}$➔ total size of the current input data size, $\mathbf{Dt_L}$➔ lowest size of dataset size

Let $\mathbf{DT_x}$➔ Number of dataset size, $\mathbf{Dt_N}$➔ total number of input modalities dataset sizes

Input= $\mathbf{Dt_{xcNc}}$, $\mathbf{Ds_{xN}}$

p(f)= $\frac{\sum_{1}^{n} \mathbf{Dt_{xcNc-}} \sum_{1}^{n} \mathbf{Dt_{xlNL}}}{\sum_{x}^{N} \mathbf{Ds_{SN}}}$

The main question of proof (3) (Equation (3)) is: What is the importance of the priority of each modality dataset input? Due to the context is unknown and the dataset is dynamic, proof (3) aims to identify the priority of each modality dataset types and number of each type.

Example#6: if the modality datasets number =2, type (x=1, 1➔ text and x=3, 3➔audio). The first dataset size is 70.000 patient's metadata of Excel sheets with 1000 cough audio of dataset size 2.

$$P(f) = = \frac{70.000-1000}{71.000} = 0.97$$

This result refers to interpret as there are 71 records for each patient. That explores the relationship between one to many.

Example#7: if the modality datasets number =2, it is applied to three modalities input for counts records metadata of patients 373 in Excel sheet and 357 in X-rays inputs as shown in results. Due to ***(Dt_{xcNc}, >= Ds_{xN})*** where

$$P(f) = \frac{373-357}{357} = 0.02$$

It interprets each patient has one X-ray that interprets the 1-1 relationship. The same problem shows different accuracy of classification disease for patients is different based on various interpretation modalities type and number. This modality priority is calculated by (Equation (3)) then applies multiplication of the temp accuracy (TempAcc) of default suitable classification results as shown in equation (13).

$$T(P(DTxn)) = P * TempAcc\ (DTxn) \qquad (13)$$

D) Context Adaption for diverse domains: refers to the reduction level filter

The domain adaption for diverse contexts can improve the object classification with reducing the uncertainty with multiple features. The computed of reduction level of domain adaption is a suggested filter to improve the object classification of the offline supervised learning that creates a model for improving decision-making as shown in proof (14) (Equation.5). It is based on various modality data types as shown in Table.10.

$$f(RL) = \sum_{Dt=1}^{n} Rweight + Mpriorty \qquad (14),\ Proof\ (4)$$

Table.10: Proof (4): (Interpretation Reduction Level)

Let f➔ fusion,

LetRL➔ Reduction Level, Dt_N➔ count data type input

Let Rweight➔ Number of dataset size, $M_{Pariorty}$ ➔ current dataset

$$\mathbf{f(RL) = \sum_{Dt=1}^{n} Rweight + Mpriorty} \qquad \mathbf{(10)}$$

The main question of proof (4) (Equation (5)) is: What is the priority of modality datasets input and the type of each one? Due to the context is unknown and the dataset is dynamic, proof (4) aims to identify the priority of each modality dataset types and number of each type.

Example#8: if the modality datasets number =2, type (x=1, 1➔ text and x=3, 3➔audio). The first dataset size is 70.000 patient's metadata of Excel sheets with 1000 cough audio of dataset size 2.

P(f (DT1,2)=70.000-1000/71.000= 0.97

P(f (DT2,3)=0

*TempDT1,1= 0.97*75% =0.72*

*TempDT2,1= 79%*0=0*

F(r(DT1,1)=7 + 0.72

F(r(DT2,1)=1+0=1

This result refers to interpret as there are 71 records for each patient. That explores the relationship between one to many.

The importance of computed reduction level which is based on the summation of weight and priority of each modality dataset is shown in changed weighted weight of neural networks as shown in Equation (6). The effected neural network with respect the multiplication of number of inputs with summation of biases of each neural network that will improve the featured of data output.

$$W=\sum_{DTxn}(\mathrm{Rl} \times \text{inputs}) + \text{bias} \qquad \textit{(6)}$$

The importance of this layer is controlling the different data topologies with interpreting heterogonous unstructured modality data input type, number, extracted relationships, weights, priority, and dataset size of each data. The contribution of this layer is shown in proofed mathematical controlling for different topologies with extracted proposed correlation with multimodality in Multicontext (unknown context). The hardness of implementing this layer is shown in management the unstructured multimodality types and characteristics with unknown Multicontext.

Layer (2): Preprocessing layer: refers to enhance the tuning data in different modality dataset types (Images, Videos, Audios, or Texts) that can be affected on weight and priority measurements. Any effect of the edited preprocessing about any data will be affected on extracted correlation measurements based on measured parameters of software-defined layer. Each modality type has pre-processing data based on normalization, cleaning, and augmentation. The importance of this layer is preprocessing automated for the different data topologies with interpreting heterogonous unstructured modality data input type. If any preprocessed data will be had a modification of the extracted correlation measurements based on measured parameters of software-defined layer.

The preprocessing layer is designed based on different four data topologies layouts (Image, Text, Audio, and Video). In addition, the preprocessing layer is designed to select the suitable deep learning technique to working parallel in the next layer.

There are two types of preprocessing that are based on the dataset size and tuning,

a) Default automated preprocessing:

Default refers to the default of different four data topologies layouts,

- Image: refers to not requires increasing number of images.
- Text: refers to the normalization, and filling missing data with null.
- Audio: refers to convert the spectrograms.
- Video: refers to split the video into images frames, compute number of frames, sorting the number of frames.

b) Inferring more dynamic training

More dynamic refers to more options of multiple topologies,

- Image: increasing number of images with augmentation, adding or removing noisy data, rotation, scaling, reflection, and crop [46], [47].

Text: cleaning, normalization, applying the trade-off between the filling missing data with null, remove missing data, remove outliers and data fill, and determine data fill and replace data fill [48], [49].

- Audio: converting spectrograms, adding noisy data, augmented data [50].
- Video: splitting video into images frames, compute number of frames, sorting the number of frames, determine the time scale of video, not requires augmentation but requires to limitation of video number of frames with respect to time that can be normalization or zeros center [51].

The preprocessed data has a changed measurement of proposed correlation based on computed weight and priority. This layer working by default properties to repair the modality dataset inputs to suitable deep learning techniques. The hardness of implementing this layer is shown in management the unstructured multimodality types and characteristics with unknown Multicontext. The outputs of this layer are processed modality datasets.

B) Fusion level two: Feature-based fusion

The second fusion level is designed based on feature-based fusion. This level relies on deep learning of different topologies into one topology of reduced matrixes. This level aims to improve the adaptive fusion with improved dempster-Shafer technique with larger filtered features and improves the accuracy evaluation. This level extracts the new proposed classification learning model between modalities dataset input into unification matrixes topologies with reduced features of all objects in the datasets. It consists of three layers that are entitled, dynamic classification layer, adaptive fusion layer, and evaluation layer. It evaluates the accuracy optimization using swarm optimizer to achieve the best fit point of accuracy.

Layer (3): A Dynamic Classification Layer: is an automated layer for improving multi-object classification and improving object detection that is based on selecting a suitable neural network concerning the input data types. A dynamic classification layer makes compatibility between the appropriate modality types and numbers in neural networks based on the input data type (image, video, audio, and Text). It is a dynamic classification that is constructed based on equation (15).

$$\{x_i, y\}^m \quad y \in \{1,2,\ 3,\ \ldots..N\} \qquad (15)$$

The dynamic deep learning layer has partial contribution in converting all modality different input topologies into one topologies of converted matrixes with reduced features. In addition, it uses the sigmoid function to extract learned features vectors from various topologies outputs numbers between zero and one, describing how much of each component should be let through. It prepares the features and objects vectors as the input of adaptive fusion layer [52], [53].

The dynamic deep learning layer has feature extraction, reduction data, converting multiple topologies into one topology of feature matrixes. The Feature extraction aims to the operation of transforming modality input datasets into the numerical data features that can be processed while preserving the information in the original data set. The main goal of feature extraction of deep learning layer is achieving better results than applying machine learning directly from multiple modality data than one modality data. The dynamic deep learning layer relies on the automated feature extraction uses specialized deep learning techniques to extract features automatically from text, images, audios, or videos without the human intervention. This technique can be very powerful to be higher results, more detailed features of data objects, quicker from raw data to developing machine learning algorithms. The used specialized deep learning techniques are AlextNet-based Image (as shown in Figure.6), LSTM -based Text (as shown in Figure.7), AlexNet-based Audio (as shown in Figure.7), GoogleNet-based Video (as shown in Figure.8) with Bilstm-based Video as shown as the following (as shown in Figure.9).

Previously, this was done through specialized feature detection, feature extraction, and feature matching algorithms. Nowadays, deep learning is very popular in image and video analysis and is known for its ability to take raw image data as input, skipping the feature extraction step. Regardless of the user's approach, computer vision applications such as image registration, object detection and classification, and content-based image retrieval require effective representation of image features. results - either implicitly through the first layers of the deep network or by explicitly applying some of the longstanding image feature extraction techniques.

A) Image: AlextNet-based Image: AlexNet is pre-trained deep neural network that is one of the best image processing for object classification in diverse contexts [54]. Reduced feature extraction of each image from diverse number of images pixels into [227*227*3]. AlexNet is designed based on 25 layers that can apply the feature extraction for image modality data represents the interesting parts of features of multiple matrixes of input images and converting into the compact feature vector as shown in Figure.4.

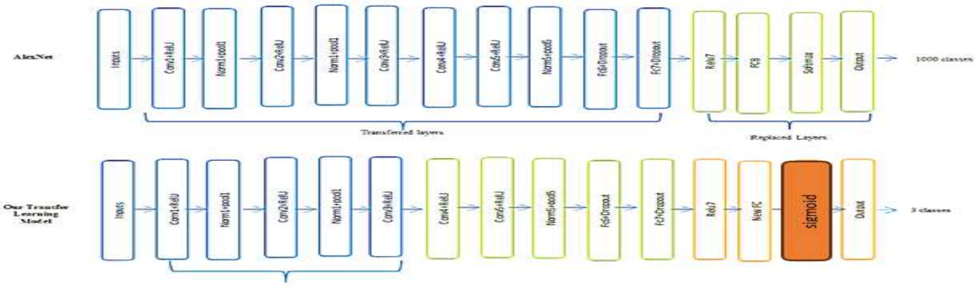
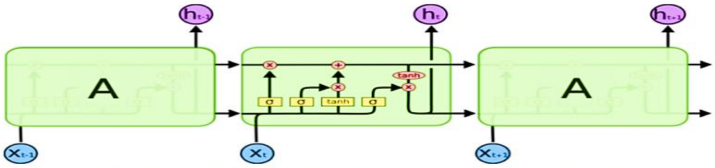


Figure.4: A Comparison between the Current AlexNet and Proposed AlexNet.

a) Text: LSTM-based Text: A Proposed LSTM depends on the feature extraction which identifies the most discriminating characteristics in texts to be more easily consume [55], [56]. A proposed LSTM is designed based on schematic process based on two matrixes, frequency features and reduced features as shown in Figure.5. First one is interpreting the feature extraction to time series data for determining the frequency of data to each feature (row) with each column. First matrixes show the importance of each feature in data. second one is the reduced features with higher effect by counting in first matrixes. Training deep learning directly with raw timeseries aims to the high data rate of information redundancy. If timeseries data has text, the proposed LSTM is running based on the Bag-of-Words it is the most widely used technique for natural language processing. During this process, they extract words or characteristics from a sentence, document, web page, etc. Then classify them according to frequency of use. So, in this whole process, feature extraction is one of the most important parts with higher frequency.



The repeating module in an LSTM contains four interacting layers.

Figure.5: Text classification for Timeseries data based on LSTM neural network

b) Audio: AlexNet-based Audio: Audio refers to a set of time-frequency transforms, including Mel spectrograms, octave and gammatone filter banks, and discrete cosine transforms (DCTs), commonly used for audio, speech, and acoustics. Spectrum of the signal using short-term Fourier transform [57]. A spectrogram shows the change in frequency content over time. The audio feature extraction is considered the most discriminating characteristics in signals, which a deep learning algorithm can more easily consume due to the high data rate and information redundancy. AlexNet is pre-trained deep neural network that is one of the best image processing for object classification in diverse contexts. Reduced feature extraction of each image from diverse number of images pixels into [227*227*3]. AlexNet is designed based on 25 layers that can apply the feature extraction for image modality data represents the interesting parts of features of multiple matrixes of input images and converting into the compact feature vector.

c) Video: GoogleNet-based Video with BiLSTM-based Video: Video is split into images' frames with sequenced images using BiLSTM. Images' frames use the GoogleNet is pre-trained deep neural network that is one of the best image processing for object classification in diverse contexts [58], [59], [60]. Reduced feature extraction of each image from diverse number of images pixels into [224*224*3] as shown in Figure.6. GoogleNet is designed based on 144 layers that can apply the feature extraction for image modality data represents the interesting parts of features of multiple matrixes of input images and converting into the compact feature vector.

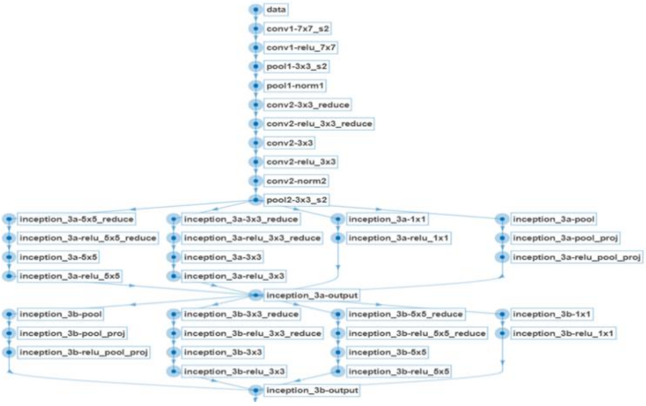


Figure.6: A part of proposed GoogLeNet

Sequenced frames of the feature extraction based on BiLSTM. BiLSTM is designed based on forward propagation and back propagation of number of frames as shown in Figure.7.

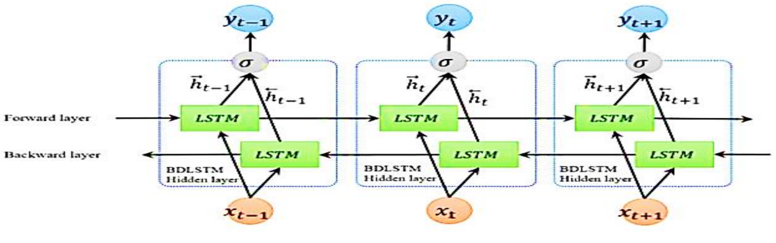


Figure.7: Video classification for sequenced text based on BiLSTM neural network.

The basic weighted CNN is designed based on weight and biases. Previous experimental example showed the convolutional neural network with a 5x5x1 input volume that interpreted into the one convolutional layer by using 5 filters. That refers to have a 2x2x1 receptive field which is relative to the one fully connected output layer with 5 neurons. The main significant question here is how many the proposed weights does the network have totally? the weight is computed based on each receptive field of a filter that refers to each filter has a single bias. This gives for an individual filter: 2*2*1+1 = 5 weights per filter. 5 filters * 5 weights = 25 weights for all filters. The conv layer presents the neural network shape (4, 4, 5).

Therefore, this produces the multiplication of (4*4*5) * 5 neurons = 400 connections. every connection has a weight. Each neuron in the dense layer also has a bias, so there are 5 additional weights. In short: 25 parameters for convective layer, 405 parameters for dense layer. A total of 430.

The Improved weighted CNN is designed based on weight and biases as shown in equation (16) $\boldsymbol{W= \sum_{DTxn} (Rl \times inputs) + bias}$ and equation (16) which It is extracted into as the following

$$w(N(Dtxn)) = \sum_{Dt=1(}^{n}(Rweight \quad + Mpriorty) \times \ inputs) + bias \quad (16)$$

$$, where\ c=bias$$

Example#9: the convolutional neural network with a Rlx5x1 input volume, followed by one convolutional layer with 5 filters.

The importance of this layer is converting the different data topologies with interpreting heterogonous unstructured modality data input into one topology based on feature-matrixes from various objects via data. The hardness of implementing this layer is shown in management the unstructured multimodality types and characteristics with unknown Multicontext. The contribution of this layer is shown in converting different topologies into one topology and changing the neural network weights based on proposed extracted correlation from diverse modality with unknown context for improving accuracy of features of each object in this layer. The outputs of this layer are two types of vectors, learned reduced vectors and non-learned vectors.

Layer (4): An Adaptive Fusion Layer: A proposed adaptive fusion layer consists of a deep learning based on Dempster-Shafer fusion model and Concatenation fusion model with reducing the redundant vectors that refer to not important features. Various modality types interpret into vectors that can measure in similar or different vectors. Complementary data discuss in various data types or characteristics. The interpretation of reduction level is interpreted. It is designed based on a tailored neural network for constructing Dempster-Shafer technique and depends on suggestions of the belief and evidence. This proposed adaptive fusion layer has benefits from enlarging the number of features and parameters in a short time-consuming. It is used for improving the classification accuracy results and prediction results. This adaptive fusion consists of the integration between the Dempster-Shafer and Concatenation fusion techniques concerning feature reduction. The Dempster–Shafer theory is one of decision fusion techniques that rely on a Frame of Discernment (FoD), which is a set of primitive hypotheses (e.g., h1 and h2) about some challenge context. The process of fusing multiple sensory data provides consistency and robustness. The proposed a proposed classifier based on Dempster-Shafer (DS) fusion and a convolutional neural network (CNN) architecture to make the set-valued classification. This classifier depends on the convolutional and pooling layers first extract high dimensional features via the input modality datasets. The features are then converted into mass functions and aggregated by Dempster's rule in a DS layer. The expected utility layer does the set-valued classification based on the mass functions. The proposed an end-to-end learning fusion model for jointly updating the network parameters. Classifier fusion is the first one, and it uses the outputs. Evidence-based decision-making is a different direction. In order to create mass functions with some frequency calibration property. Designing evidential classifiers, an evidential classifier's outputs can be applied to decision-making. The results of an are made possible by the generality and expressiveness of the DS formalism. Evidence-based classifiers offer more information than traditional classifiers. a SoftMax layer-enhanced neural network that can transform an input feature vector into a

probability distribution or any other distribution. The outcomes of an evidential classifier can be utilized for decision-making. Much appreciated to the simplification and expressiveness of the DS formalism, the yields of an evidential classifier give more data than those of ordinary classifiers (e.g., a neural network with a SoftMax layer) that convert an input include vector into a likelihood distribution or any other distribution.

Adaptive fusion layer is designed as a tailored weighted neural network including belief, evidence, and plausibility. It is designed based on two fusion statistical methods of Dempster-Shafer and Concatenation fusion to improve classification accuracy. It improves the uncertainty data in this model is given by Belief and Plausibility. Belief will lead to believe in some possibility by bringing out some evidence. Plausibility will make evidence compatible with possible outcomes. The importance of Dempster-Shafer theory DST is an evidence theory; it combines all possible outcomes of the problem. Hence it is used to solve problems where there may be a chance that different evidence will lead to some different result. That differs from Bayesian theory is only concerned about single evidence. Bayesian probability cannot describe ignorance.

Each sensor, sensor Si, for instance, will participate in it notice by specifying its beliefs over Θ. The function is known the "probability mass function" of the sensor Si, indicated by mi. So, according to sensor Si's notice, the probability that "the detected person is user A" is indicated by a "confidence interval".

The adaptive fusion approach is designed based on making two fusions on two fusion models with improved the decision fusion high level with some detailed features of low fusion level. It makes parallel fusions of the Dempster-Shafer and Concatenation then extracts the not important features and reduces these features. The adaptive fusion layer refers to drawing the full vision of the modal's classification. That provides the unification target of multiple sensory data classifications in various smart environment systems. The present work implements the majority voting of the different CNN-based pre-trained models with an adaptive fusion technique.

The proposed mathematical fusion structure is designed into two classification levels to make the aggregation weighted evidence that constructs a new tailored neural network for two fusion models, Dempster-Shafer fusion and Concatenation fusion working the parallel for improving the classification accuracy results. The MultiFusion learning model has a classifier that is designed based on parallel working:

a) Evidential Dempster-Shafer neural network with automated computed the belief and evidence; and the output of classifier.

It is constructed based on a tailored neural network for constructing Dempster-Shafer technique and depends on suggestions of the belief and evidence. This proposed adaptive fusion layer has benefits from enlarging the number of features and parameters in a short time-consuming. It is used for improving the classification accuracy results and prediction results. The Dempster–Shafer theory is one of decision fusion techniques that rely on a Frame of Discernment (FoD), having primitive hypotheses group about some context problems [61]. The process of fusing multiple sensory data provides consistency and robustness.

The proposed classifier based on Dempster-Shafer (DS) fusion and Convolutional Neural Network (CNN) architecture to make worthwhile classification. This classifier depends on the convolutional and pooling layers first extract high dimensional features via the input modality datasets. The features are then transformed to mass features and summed in the Dempster-Shafer fusion layer using Dempster's rule. The predictable utility layer does the group-valued classification based on the mass functions. The proposed an end-to-end learning fusion model for jointly updating the network parameters. Classifier fusion is the first one, and it uses the outputs. Evidence-based decision making is considered as another direction. In order to create mass functions with some frequency calibration property. Designing evidential classifiers, an evidential classifier's outputs can be applied to decision-making. The results of an are made possible by the generality and expressiveness of the DS formalism. Evidence-based classifiers offer more information than traditional classifiers.

Adaptive fusion layer is designed as a tailored weighted neural network including belief, evidence, and plausibility. It is designed based on two fusion statistical methods of Dempster-Shafer and Concatenation fusion to improve classification accuracy. Dempster-Shafer reduces the uncertainty data in this model is given by Belief and Plausibility. Belief refers to the believe in some possibility by bringing out some evidence. Plausibility makes evidence consistent with potential outcomes. The importance of Dempster-Shafer theory DST is an evidence theory; it gathers all possible results of the problem. Hence it is utilized to proposed a solution of problems. That differs from Bayesian theory is only concerned about single evidence. Bayesian probability does not allow to depict the ignorance [62]. Each sensor, sensor Si, for instance, will participate in it notice by specifying its beliefs over Θ. The function is known the "probability mass function" of the sensor Si, indicated by mi. So, it takes care of the sensor Si's notice, the probability that "the detected person is user A" is indicated by a "confidence interval" as illustrated in equation (17).

$$[Belief, (A), Plausibility_{\,i}\,(A)], \qquad (17)$$

The lower bound of the confidence interval is the belief confidence as illustrated in equation (18) and equation (19), that accounts for all evidence Ek that supports the given proposition "user A". The plausibility confidence refers to the upper level of the confidence interval, and it can compute all the notices that do not revoke by the given proposition.

$$Belief_i\,(A) = \textstyle\sum_{E.\subset A} m(E_k), \qquad (18)$$

$$Plausibility_i\,(A) = 1 - \textstyle\sum_{E.\sqcap A=\emptyset} m(E_k), \qquad (19)$$

The three most common decision rules are maximum confidence, maximum belief, and maximum belief without overlapping belief intervals. For each conceivable proposition (e.g. user-A), the Dempster-Shafer theory allows a unified rule for the opinion mi of sensor Si and the opinion mj of Sj as shown in equation (20). This merge rule is applied repeatedly to be more general: if the project proposes to consider mj not as sensor Sj but as the collected Dempster-Shafer fusion rule of sensor Sk and sensor Sl:

$$\left(m_i \oplus m_j\right)(A) = \frac{\sum_{E_i\cap E_j=\emptyset} m_i(E_k)m_j(E_k)}{\sum_{E_i\cap E_j=\emptyset} m_i(E_k)m_j(E_k)}, \qquad (20)$$

The proposed Convolutional Neural Network (CNN) for deep learning classification demonstrates how to Load modality input data, Determine the neural network architecture, Specify the neural network training properties, and Train the selective deep learning network as the following.

- Image Input Layer: The image input layer is the size of the image, in this case 28 x 28 x 1. These numbers correspond to the height, width, and channel size. Digital data includes grayscale images, so the channel size (color channel) is 1. For color images, the channel size is 3, corresponding to the RGB value. Training the network can automatically shuffle the data at the start of each epoch during training.

- Convolutional Layer: In the convolution layer, the first argument is the filter size, which is the height and width of the filter that the training function uses when analyzing the image. In this example, the number 3 represents the filter size of 3 x 3. We can specify different sizes for the filter height and width. The second argument is the number of filters, numFilters, corresponding to the number of neurons connected to the same region of the input. This setting determines the number of feature tags. Use the name-value pair “Padding” to add padding to the input feature map. For a convolutional layer with a default step of 1, the "same" padding ensures that the spatial output size is the same as the input size.

- Batch Normalization Layer: Batch normalization layers normalize the activations and gradients propagated through the network, making training the network a simpler optimization problem. Use batch normalization layers between convolutional and nonlinear layers, such as ReLU layers, to speed up network training and reduce network initialization sensitivity.

- ReLU Layer: The batch normalization layer is followed by a nonlinear activation function. The most common activation function is the rectified linear unit (ReLU).

- Max Pooling Layer: Convolutional layers (with activation functions) are sometimes followed by a down sampling operation that reduces the spatial dimension of the feature map and removes redundant spatial information. Down sampling allows you to increase the number of filters in deeper convolutional layers without increasing the amount of computation required for each layer. The max pooling layer returns the maximum value of the input's rectangular regions, specified by the first argument, pool size. In this example, the size of the rectangular area is [2,2]. The argument to the name-value pair "Stride" specifies the step size that the training function takes when scanning along the input.

- Fully Connected Layer: The convolution and down sampling layers are followed by one or more fully connected layers. As the name suggests, a fully connected layer is a layer in which neurons connect to all neurons in the previous layer. This layer combines all the features learned from previous layers on the image to identify larger patterns. The final fully connected layer combines features to classify images. Therefore, the Output Size parameter in the final fully connected layer is equal to the number of layers in the target data. In this example, the output size is 10, which corresponds to 10 classes. The output is fused vector as shown in equation (21),

$$D[\ ] = \{x1, x2,...., xn\}; \qquad (21)$$

b) Concatenated neural network to classify the characteristics estimation for multimodal was collected by multiple sources/sensors in offline mode; and the output of classifier as shown in equation (22). The output is fused vector.

$$W[\ ] = \{y1, y2,...., yn\}; \qquad (22)$$

The main characteristics of Concatenation fusion have greater number of features and not important to learned the data to apply the concatenate fusion from various datasets. The major advantages of concatenation fusion are shown in adding more information, uncertainty interval reduces. The main disadvantage has not important data.

The Filter classifier is designed to filter classification

c) The subtraction of the redundant vector of feature classes detected in the first vector by the classes detected on the second vector, which is equivalent to the initialization of our data fusion algorithm the output is Fused vector of F []. The interpretation of reduction level is interpreted by the weight of various relationships between parameters (Pweight) and the priority values between input modality types (Mpriority) as proven in equation (23) as proven in Proof (4). The relationships discuss parameters and their relationships. The reduction level is based on similar vectors or different vectors. Data Reduction relies on the parameters and their relationships or conditions between themselves and the priority between modality data inputs.

- **SoftMax Layer:** The SoftMax activation function normalizes the output of the fully connected layer. The output of the SoftMax layer consists of positive numbers that sum to 1, which can then be used by the classification layer as the classification probability.

- **Classification Layer:** The last layer is the classification layer. This layer uses the probability returned by the SoftMax activation function for each input to assign the input to one of the mutually exclusive classes and calculate the loss.

The Output includes the fused feature vector from the two inputs as shown in equation (23). The feature fusion includes the Input,

$$F[\ ] = \{y1, y2, \ldots, yn\}; \qquad (23)$$

The deep neural network is described working as the following a) the weighted sum of the inputs is calculated. b) the bias is added. c) the result is fed to an activation function. d) specific neuron is activated.

The improvement of Dempster-Shafer is shown in automated neural network and getting larger number of features. The improvement of Dempster-Shafer aims to classify multi-label classification. The adaptive fusion approach is designed based on making two fusions on two different levels, high and low. It makes parallel fusions of the Dempster-Shafer and Concatenation then extracts the not important features and reduces these features.

The contribution of this layer is shown in improved computational and extracted features of dempster-Shafer fusion by concatenation fusion to improve the fused multi object classification from diverse multimodality in Multicontext or unknown context. The adaptive fusion layer refers to drawing the full vision of the modal's classification. That provides the unification target of multiple sensory data classifications in various smart environment systems. The present work executes the majority voting of the diverse CNN-based pre-trained models with an adaptive fusion technique. Much appreciated to the simplification and expressiveness of the DS formalism, the yields of an evidential classifier give more extracted data than the ordinary classifiers (e.g., a neural network with a sigmoid function to extract vectors layer) that transform an input include vector into a likelihood distribution or any other distribution. A sigmoid function is considered a support of multi-label object classification.

The importance of this layer is converting the different vectors from various matrixes of topologies layouts. The hardness of implementing this layer is shown in management the unstructured multimodality types and characteristics with unknown Multicontext. The output featured filtered vector.

Layer (5): Evaluation layer: relies on two parts, Part one is evaluating the training accuracy and optimization results of multiple smart systems. It improves accuracy results 96% to 98% in various contexts. The experiments are applied to various multimodal inputs for diverse contexts which have the common factors of smart systems as the following, smart military and smart health. It measures the accuracy and optimization results in multiple smart context systems. This layer classifies data into two types, training data and testing data. It measures the accuracy, precision, recall, F1-measures [63]. The training applies particle swarm optimization to improve the accuracy evaluation. The training changes the hyper parameter in 30 times to achieve the best fit point.

The importance of this layer is training of data to achieve the best fit accuracy point. The hardness of implementing this layer is shown in management the unstructured multimodality types and characteristics with unknown Multicontext. The contribution of this layer is shown in achieving the best accuracy result with changing hyperparameters to achieve to the best optimized point. The output featured filtered vector. Multi-class deep learning model is designed to the neural network can make a prediction analysis of multi-class. it computes the confidence in c of the SoftMax.

3.4 Mathematical Formulation Model:

This section describes the mathematical formulation model which is a proof of the proposed Adaptive Multimodal Fusion Framework in Contextual Representation based on Late Fusion Level using MultiFusion Learning Model and Improved Evidential Deep Learning Dempster-Shafer as shown in Table.11.

Table.11: A proposed Mathematical formulation Model based on Multimodal fusion framework.

	Input: *select number of modality dataset input, and modality type.* *Then open the path of each modality dataset folder.*
	Output: *construction of fused filter features vector relative to multi-object classified vectors dynamically.* *Classification accuracy results to multi-object classification in Multicontext*
	Modality input: *count the total number of modality dataset inputs.*
	Initialization of Variables *(Declarations)*
	Let DT➔ data type, let n= the total number of modality inputs, *Let* $\boldsymbol{Dt_x}$➔ *a type of degree of each data,* $\boldsymbol{Dt_N}$➔ *total number of inputs of each type* *Input= DT#, DTx*

```
Let $DT_{x1}$➔ Image data type, $DT_{x2}$➔ Audio Data type, $DT_{x3}$➔ Text data type, $DT_{x4}$➔ Video
Let DT➔ data type, let n= the total number of modality inputs,
Let $Ds_x$➔ total size of neighbor biggest datasets,
Let $Ds_{xc}$➔ Number of dataset size, $Dt_{Nc}$➔ current dataset
Relationships (1-1, 1-m, m-m)
Let $DT_c$➔ total size of the current input data size, $Dt_L$  lowest size of dataset size

Let $DT_x$➔ Number of dataset size, $Dt_N$➔ total number of input modalities dataset sizes
Let f➔ fusion,

LetRL➔ Reduction Level, $Dt_N$➔ count data type input
Let Rweight➔ Number of dataset size, $M_{Pariorty}$ ➔ current dataset
WNN=0
$\{x_i, y\}^m \;\; y \in \{1, 2, 3, \;\ldots\ldots N\}$
c=bias
Layer 1  Software defined fusion layer
while (i< modality-folders) do// reach each utterance one by one extraction
{
Start
For (DTx=1; i<n; x++)

For (DTi=1; i<n; i++)

asure Multimodality Adaption
    Compute the Modality Dataset number for each modality input
    Compute the Modality Dataset size for each modality input
    Compute Modality dataset Type for each modality input
        $I(n) = \sum_{i=1}^{N} \sum_{j=1}^{x} Dt_{xN}$
ract Multimodality Relationships
    Compute the Modality dataset Weight for each data input
        $I(w)= \frac{\sum_1^n Dt_{x1N1}}{\sum_x^N Ds_{xN}}$

ltimodality Priority
    Compute each Modality dataset Priority
        $p(f)= \frac{\sum_1^n Dt_{xcNc-} \; \sum_1^n Dt_{xlNL}}{\sum_x^N Ds_{SN}}$

    then,
    Compute the temp of total measurement or priority
        T(P(DTxn)) =P * TempAcc (DTxn)
ıtext Adaption
    Measure the Modality Reduction Level for each data
        $f(RL) = \sum_{Dt=1}^{n} Rweight \;\; + Mpriorty$

    Compute value of each data.

Layer 2  Preprocessing layer
Processed data based on computed relationships between data
For (int i=0; DTix; i<ni++)   x➔type


If (DTs1 =< DTsL) //[fused data in the relationship interval limit [1-1/10]



Print (‘Not required to change size or larger number);

Else {
l

Return layer 1; //change weight, priority
For (DTi=1; i<n; i++)

// Select the deep learning techniques working parallel
if (i==1):

ad Augmentation; //load crop; //load rotation; load Kalman filter to check noisy

Else if (i==2):

l filling missing data, normalization and cleaning;

Else if (i==3):

t ('Audio');
    //convert audio into spectrograms
    //load add noisy on it
```

	Else (i==4):
	
	//load normalization
	//load zero center;

Layer 3	***learning layer***
	Construct apply suitable selective deep learning techniques based on number and type of inputs working parallel ***For (DTx=1; i<n; x++)*** ***For (DTi=1; i<n; i++)*** *ect the deep learning techniques working parallel* *if (i==1):* *Load AlexNet* *Create Reduced images input into matrixes (not-learned)* *Create Reduced images input into matrixes (learned)* *Apply the reduction level with changing in applied neural networks* *Change the new the weighted neural network of each modality type* *WNN=* $\sum_{DTxn}$ ***(Rl ×inputs) + bias*** Which It is extracted into as the following *WNN(N(Dtxn)) =* $\sum_{Dt=1(}^{n}(\boldsymbol{Rweight} + \boldsymbol{Mpriorty})\times \boldsymbol{inputs}) + \boldsymbol{bias}$ *Create Reduced not learned Feature Vector* *Create Reduced learned Feature Vector* *Else if (i==2):* *//create LSTM* *//create frequency of effected features* *Create Reduced images input into matrixes (not-learned)* *Create Reduced images input into matrixes (learned)* *Apply the reduction level with changing in applied neural networks* *Change the new the weighted neural network of each modality type* *WNN=* $\sum_{DTxn}$ ***(Rl ×inputs) + bias*** Which It is extracted into as the following *WNN(N(Dtxn)) =* $\sum_{Dt=1(}^{n}(\boldsymbol{Rweight} + \boldsymbol{Mpriorty})\times \boldsymbol{inputs}) + \boldsymbol{bias}$ *Create Reduced not learned Feature Vector* *Create Reduced learned Feature Vector* *Else if (i==3):* *t ('Audio');* *//convert audio into spectrograms* *//load Alexnet* *Create Reduced images input into matrixes (not-learned)* *Create Reduced images input into matrixes (learned)* *Apply the reduction level with changing in applied neural networks* *Change the new the weighted neural network of each modality type* *WNN=* $\sum_{DTxn}$ ***(Rl ×inputs) + bias*** Which It is extracted into as the following *WNN(N(Dtxn)) =* $\sum_{Dt=1(}^{n}(\boldsymbol{Rweight} + \boldsymbol{Mpriorty})\times \boldsymbol{inputs}) + \boldsymbol{bias}$ *Create Reduced not learned Feature Vector* *Create Reduced learned Feature Vector* *Else (i==4):* *//load GoogLeNet* *//Create BiLSTM* *Create Reduced images input into matrixes (not-learned)* *Create Reduced images input into matrixes (learned)* *Apply the reduction level with changing in applied neural networks* *Change the new the weighted neural network of each modality type* *WNN=* $\sum_{DTxn}$ ***(Rl ×inputs) + bias*** Which It is extracted into as the following *WNN(N(Dtxn)) =* $\sum_{Dt=1(}^{n}(\boldsymbol{Rweight} + \boldsymbol{Mpriorty})\times \boldsymbol{inputs}) + \boldsymbol{bias}$ *Create Reduced not learned Feature Vector* *Create Reduced learned Feature Vector*
Layer4:	***Adaptive Fusion layer***

```
//Dempster-Shafer fusion
            Set
Compute belief and plausibility given $\Omega$ and $Fm(\Omega)$
// given the mass assignments as assigned by the detectives
Initialize an array Bel with features, BEL =0, for each node
            $A \epsilon 2^{\Omega}$ ;
Initialize an array pl with one element Pl(A)=0 for each node $A \epsilon 2^{\Omega}$ ;
For each $A \epsilon 2^{\Omega}$ do
For each $S \epsilon Fm(\Omega \ )$ do
If ( $S \cap A$) $\neq \emptyset$
 $Plausibility_i\ (A) = 1 - \sum_{E.\sqcap A=\emptyset} m(E_k),$
Pl(A)= Pl(A)+m(S);
            End

If (S $\subsetneq$A ) then
Pl(A)= Pl(A)+m(S);
$Belief_i\ (A) = \sum_{E.\subset A} m(E_k),$
            End
    //create two arrays of belief and plausibility to support the fusion of number of features
    of all objects
$(m_i \oplus m_j)(A) = \frac{\sum_{E_i \cap E_j=\emptyset} m_i(E_k) m_j(E_k)}{\sum_{E_i \cap E_j=\emptyset} m_i(E_k) m_j(E_k)},$
The Dempster-Shafer neural network output is the fused classified vector as shown in equation,

                                                      D[ ] = {x1, x2,...., xn};
    End
    //Concatenation fusion
Concatenated neural network output is the fused classified vector output of classifier
     W[ ] = {y1, y2,...., yn};
     //Proposed NN-based Filter Classifier
Remove identical feature vectors
Output featured fused vector with relevant number of objects.

Layer 5   Evaluation layer
 Accuracy Evaluation
For (T=0; T<30; T++)

Training Accuracy
 $F - measure = \frac{2\ Recall \cdot Precsision}{Recall+Precision}$
Particle Swarm Optimization (threshold)
Change the hyperparameters to achieve the best fit point
$v_i^{t+1} = v_i^t + c_1 U_1^t + (pb_i^t - p_i^t) + c_2 U_2^t + (gb^t - p_i^t)$
Accuracy Output
}
   }
End
```

3.4 MultiFusion Models for Same Modality Input based on Deep Neural Networks and Dempster-Shafer Theory

The proposed MultiFusion model can be adaptive and applicable for the same multimodality (Image, Text, Audio, and Video) and the different multimodality. Proposed dynamic classification for multimodality can be affected on fusion model as the following, proposed AlexNet-based Image Fusion, proposed LSTM-based Text Fusion, proposed AlexNet-based Audio Fusion, and GoogleNet-BiLSTM based Video Fusion.

Multi-modal fusion becomes one of the recent research trends in artificial intelligence. Multi-modal fusion benefits the complementarity of heterogeneous data and provides a reliable classification for the model [48]. It changes over information from different single-mode representations to a compact multi-modal representation. The multi-modal combination is a greatly vital inquiry about direction and core technology in the multi-modal field inquire about. The multi-modal combination utilizes the complementary data show in multi-modal information by fusing diverse modalities. It integrates the context information (location, time, and camera parameters), and content information with the domain-oriented semantic ontology. One of the problems of multi-modal fusion is expanding inference to multi-modal while keeping the demonstration and calculation complexity sensible.

5.5.1 The proposed AlexNet-based Image Fusion:

This is a proposed model for classifying the same Image modality input with diverse characteristics, features, goals, and dataset size. Each image modality input requires making feature extraction and reducing the image size in (227×227×3) and managed in groups as AlexNet is organized in 25 layer. The classification layer includes Convolution, normalization, Max-Pooling layer, fully connected layers. The output of this layer is more feature data, that has 1000 dimensions, so it reaches better retrieval results as shown in Figure.8.

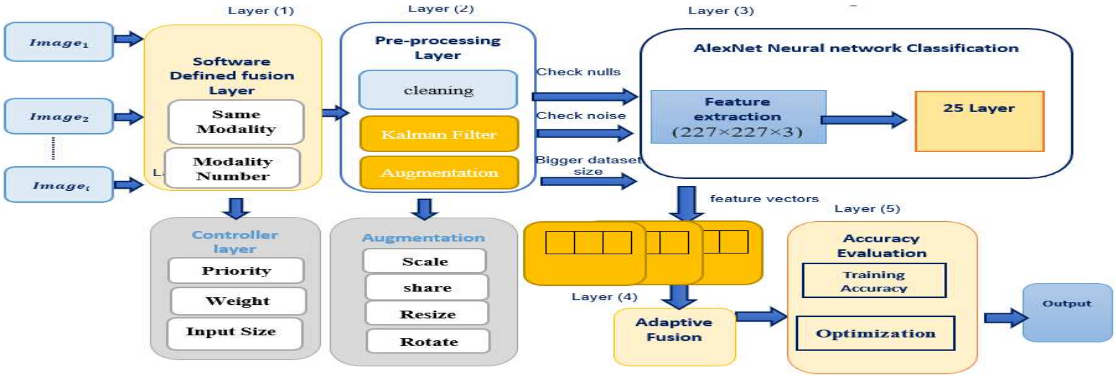


Figure.8: The proposed AlexNet based Image Fusion

5.5.2 The proposed LSTM-based Text Fusion

Long-short term memory (LSTM) is very useful for sequenced data based on time as text classification. LSTM is powerful in sentence segmentation, word segmentation, text normalization, part-of-speech tagging, and pronunciation in text analysis and speech synthesis. The proposed of the LSTM is to decide what information we discard from the state of the cell. This decision is made by a sigmoid layer entitled the "forgotten gate layer". It observes ht−1ht−1 and xtxt, and outputs a number between 0 and 1 for each number in the cell state Ct−1Ct−1. LSTM interprets the sentences input and the excel sheets. The proposed LSTM architecture (as shown in Figure.9) is constructed based on equation (24), equation (25), and equation (26) as the following steps:

- **Step 1:** Compute the sum of all the inputs (x) according to their weights and include the bias term.

$$Z = (weights\ * X) + bias \qquad (24)$$

- **Step 2:** Estimate an activation function to calculate the expected output

$$Y = Activiation\ (Z) \qquad (25)$$

- Steps 1 and 2 are performed at each layer, if we recollect, this is nothing but forward propagation! No activation functions. Our equation for y essentially becomes

$$Y = Z = (Weights\ * X) + bias \qquad (26)$$

The basic concept of LSTM is the state of a cell and its diverse gates. Sigmoid: Gates includes sigmoid activations. A sigmoid activation is considered very similar to the Tanh activation as mention in equation (27). Instead of squishing values between -1 and 1, with respect to the interval values between 0 and 1.

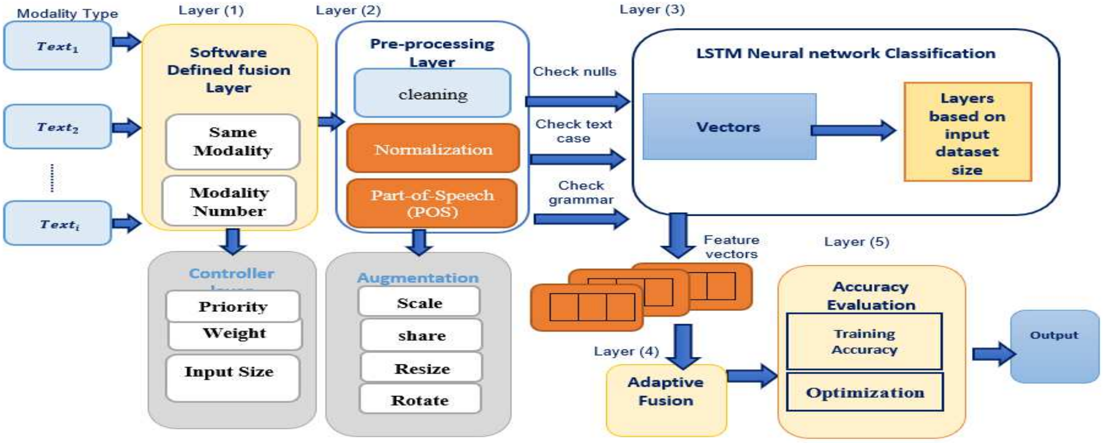


Figure.9: The proposed LSTM based Text Fusion

5.5.3 The proposed AlexNet-based Audio Fusion

Audio feature extraction converts to classify the audio into spectrogram images. For Audio (Signal features and time-frequency transformations), Feature extraction defines the most discriminating properties in signals. Audio interpretation deals as image input and uses transfer learning techniques that is more easily consumed as shown in Figure.10.

$$b^{i}_{x,y} = a^{i}_{x,y} / \left(k + \alpha \sum_{j=\max(0,i-n/2)}^{\min(N-1,i+n/2)} (a^{j}_{x,y})^2 \right)^{\beta} \quad (27)$$

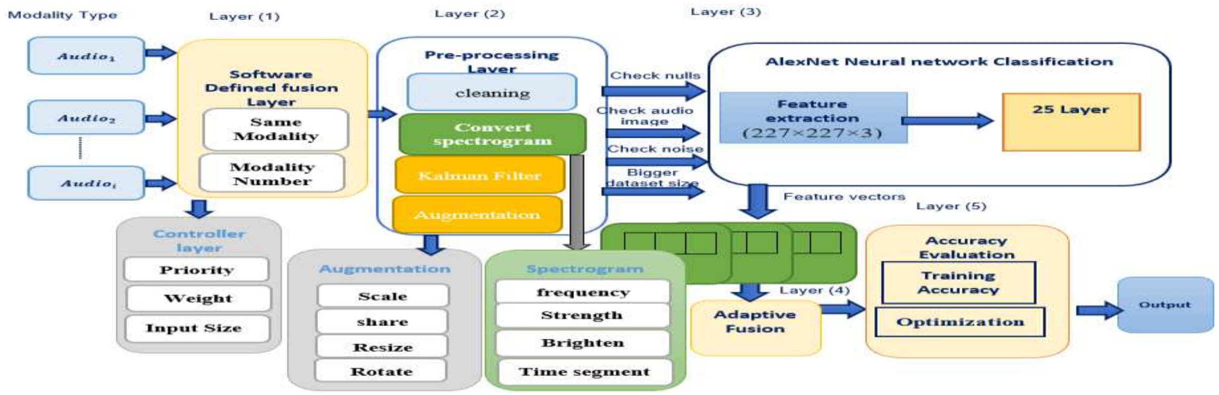


Figure.10: The proposed AlexNet based Audio Fusion.

5.5.4 The proposed GoogLeNet and BiLSTM based Video Fusion

Video classification interprets a series of images that is constructed based on deep learning techniques that handle video classification as performing image classification a total of N times, where N is the total number of frames in a video as shown in Figure.11. Video feature extraction interperts the video frames into image frames and video sequance (224×224×3). Video classification contains four steps, prepare training data, select a video classifier, train and evaluate the classifier, and utilizing the classifier to process video data. Video classification interprets video input splits into video frames and audio frames.

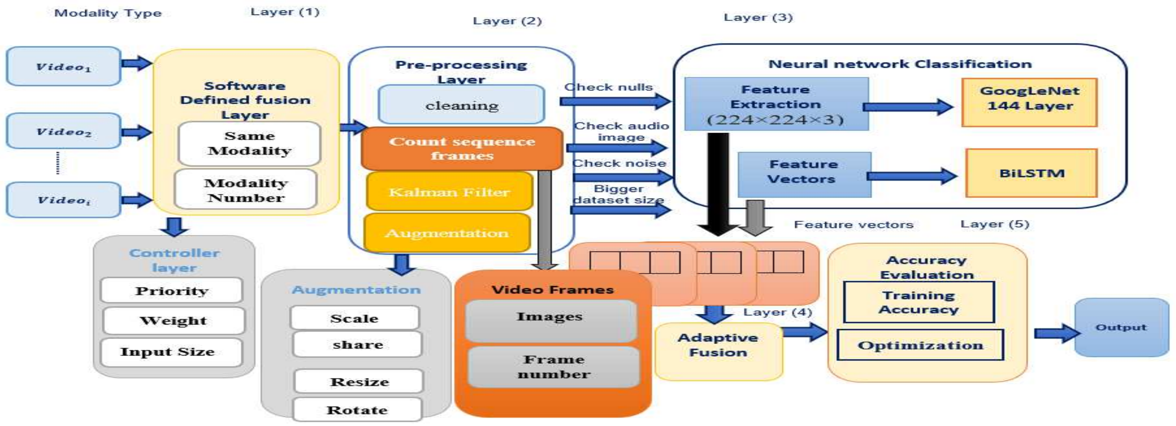


Figure.11: The proposed GoogleNet and BiLSTM based Video Fusion.

A) GoogLeNet: is a kind of convolutional neural network depending on the Inception architecture. It utilizes the Inception modules, that support the network to select between multiple convolutional filter sizes in each block. GoogLeNet is trained to process over a million images and can classify up to 1000 images object classes (such as keyboard, coffee mug, pencil, and many animals) as shown in Figure.14. The network learned multiple features for many different images. Using sigmoid activation function presents in equation (28), It will return vector of features of all objects in the modality dataset. The structure of the proposed GoogLeNet and basic GoogLeNet neural network.

$$\boldsymbol{y(i)} = \frac{exp(a^{(j,\theta)})}{\sum_{j=1}^{N} exp(a^{(j,\theta)})} \quad (28)$$

B) **BiLSTM:** The video classification converts videos to sequences into feature vectors using a pre-trained convolutional neural network (CNN), such as GoogLeNet, to extract the vector features via individual frame separately. It makes training an LSTM network on the sequences to forecast the video labels. The video classification here is used biLSTM function to check the sequence in the two directions to achieve the best accuracy. The BiLSTM is constructed based on 50 epochs, 2000 hidden layers. Video classification for sequenced text based on BiLSTM neural network and the example output is shown in Figure.11.

3.5 MultiFusion Model for Different Modality Input based on Deep Neural Networks and Dempster-Shafer Theory

The proposed adaptive framework supports the MultiFusion models for different modality input. But the MultiFusion model for different modality can interpret bimodal, trimodal, and a greater number of modality inputs with different data types as shown in Figure.12. The hardness of this model is how to apply the suitable pre-processing techniques, feature extraction techniques, deep learning techniques parallel based on modality type and number.

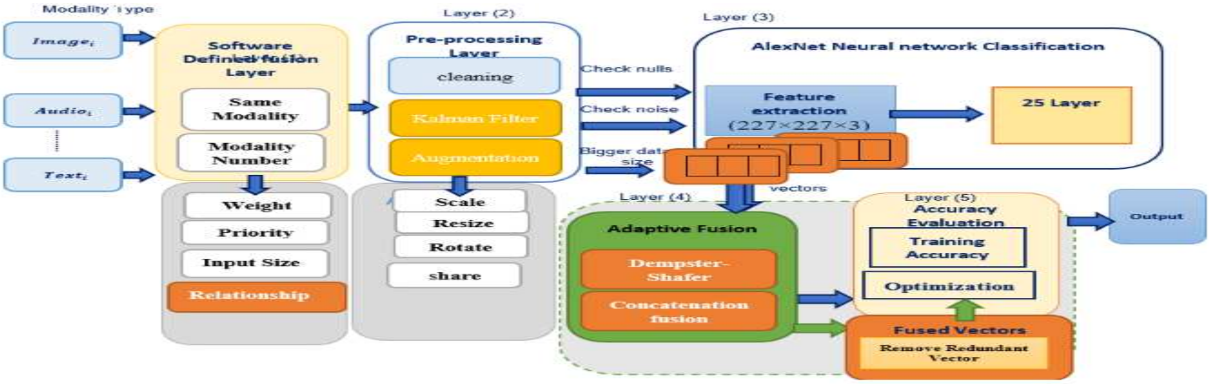


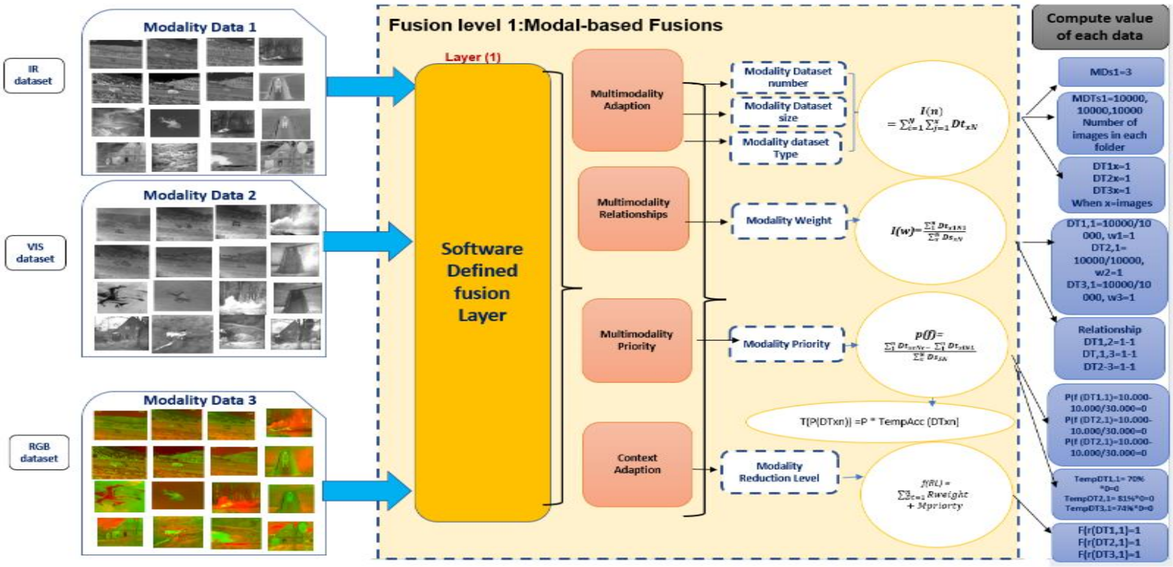
Figure.12: Multi-Fusion Models for Different Modality Input.

3.6 Explainable Examples

This section applies two full explainable examples for same modality types and different modality dataset types. The first same modality types include the five layers that are shown in Figure.15, Figure.16, Figure.17, Figure.18, Figure.19, and Figure.20. This second different modality includes the five layers that are shown in Figure.21, Figure.22, Figure.23, Figure.24, Figure.25, Figure.26, Figure.27, Figure.28 and Figure.29.

3.6.1 Explainable Example 1: (Same Modality Types: three Images)

The first full explainable example for the same modality input types is designed and applied based on the five layers that are shown in Figure.13, Figure.14, Figure.15, Figure.16, Figure.17, Figure.18. Figure.13 shows Explainable Example 1 (Layer 1): Software-Defined Fusion Layer. Figure.14 shows Explainable Example 1 (Layer 2): Preprocessing Layer. Figure.15 shows Explainable Example 1 (Layer 3): Deep Learning Layer. Figure.16


IR dataset
VIS dataset
RGB dataset
Modality Data 1
Modality Data 2
Modality Data 3
Fusion level 1:Modal-based Fusions
Layer (1)
Software Defined fusion Layer
Multimodality Adaption
Multimodality Relationships
Multimodality Priority
Context Adaption
Modality Dataset number
Modality Dataset size
Modality dataset Type
Modality Weight
Modality Priority
Modality Reduction Level
$I(n) = \sum_{i=1}^{N}\sum_{j=1}^{x} Dt_{xN}$
$I(w) = \frac{\sum_1^n Dt_{x1N1}}{\sum_1^n Ds_{xN}}$
$p(f) = \frac{\sum_1^n Dt_{xrNr} - \sum_1^n Dt_{x1N1}}{\sum_1^n Ds_{xN}}$
T[P(DTxn)] =P * TempAcc (DTxn]
$f(RL) = \sum_{c=1}^{n} Rweight + Mpriorty$
Compute value of each data
MDs1=3
MDTs1=10000, 10000,10000 Number of images in each folder
DT1x=1
DT2x=1
DT3x=1
When x=images
DT1,1=10000/10000, w1=1
DT2,1=10000/10000, w2=1
DT3,1=10000/10000, w3=1
Relationship
DT1,2=1-1
DT,1,3=1-1
DT2-3=1-1
P(f (DT1,1)=10.000-10.000/30.000=0
P(f (DT2,1)=10.000-10.000/30.000=0
P(f (DT2,1)=10.000-10.000/30.000=0
TempDT1,1= 70% *0=0
TempDT2,1= 81%*0=0
TempDT3,1=74%*0=0
F[r(DT1,1]=1
F[r(DT2,1]=1
F[r(DT3,1]=1

shows Explainable Example 1 (Layer 4): Adaptive Fusion Layer. Figure.17 presents Explainable Example 1 (Layer 4): Adaptive Fusion Layer. Figure.18 shows Explainable Example 1 (Layer 4): Adaptive Fusion Layer. And Figure.19 shows Explainable Example 1 (Layer 5): Evaluation Layer.

Figure.13: Explainable Example 1 (Layer 1): Software-Defined Fusion Layer.

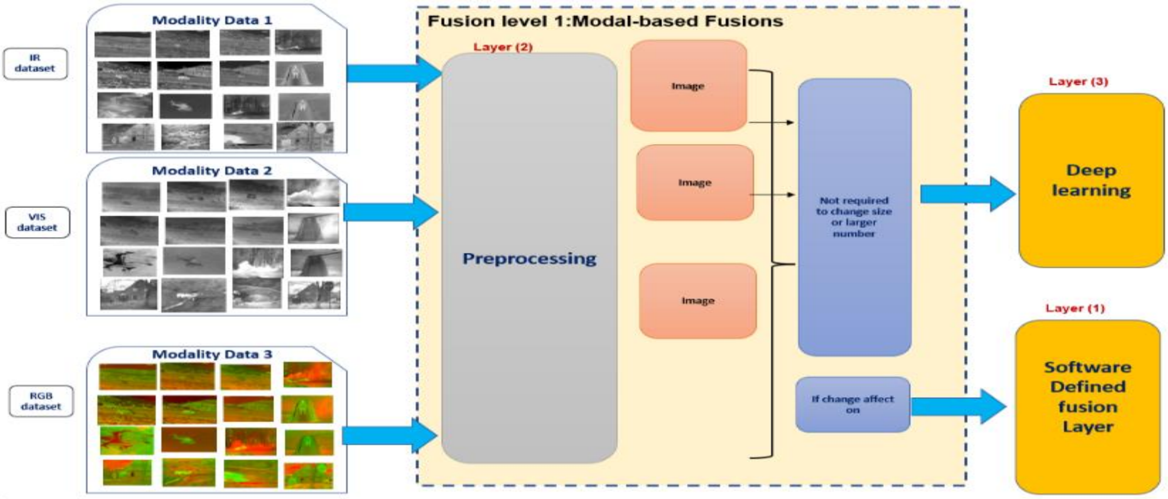


Figure.14: Explainable Example 1 (Layer 2): Preprocessing Layer.

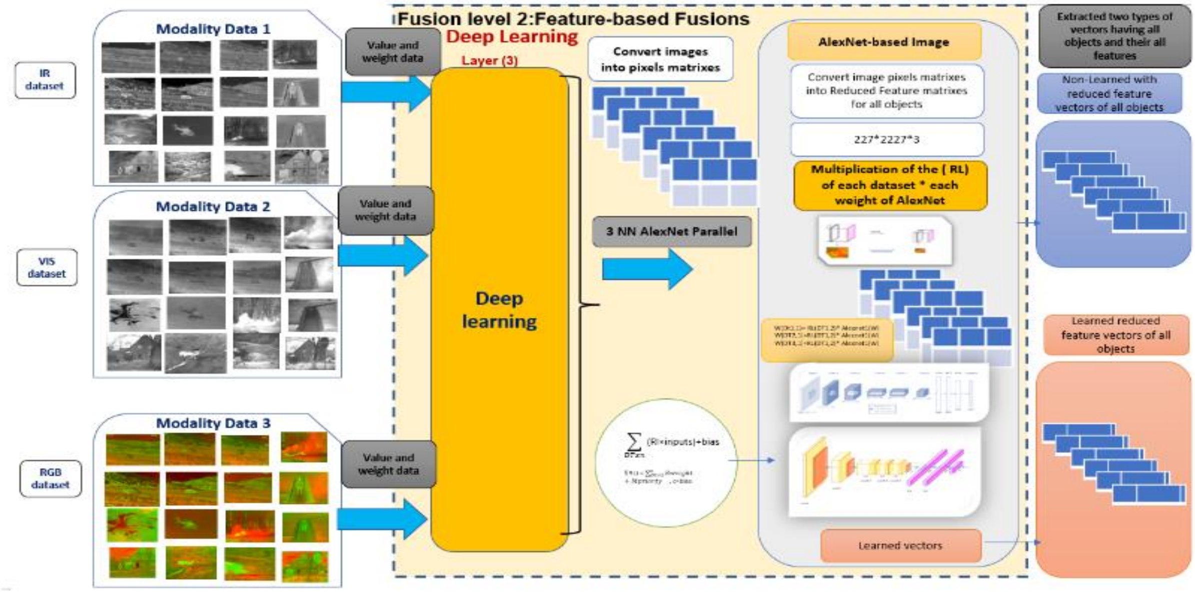
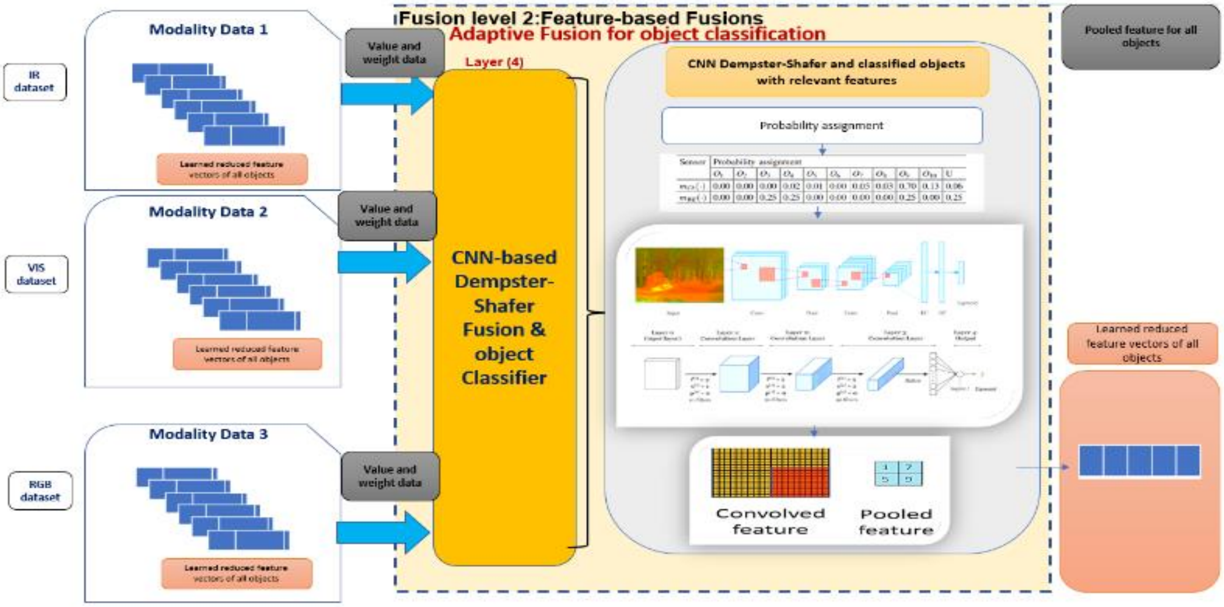
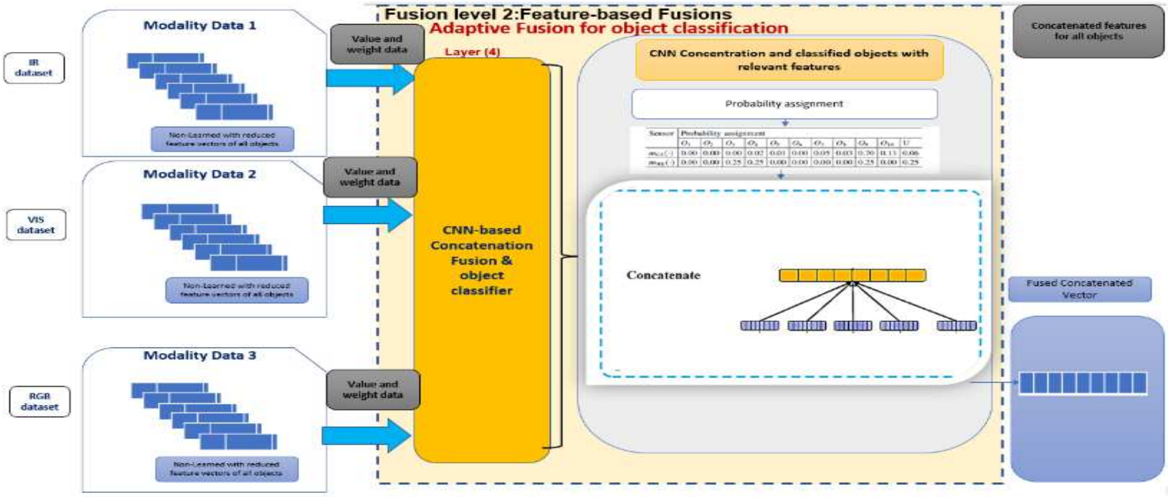

Modality Data 1
IR dataset
Modality Data 2
VIS dataset
Modality Data 3
RGB dataset
Value and weight data
Value and weight data
Value and weight data
Fusion level 2:Feature-based Fusions
Deep Learning
Layer (3)
Deep learning
Convert images into pixels matrixes
3 NN AlexNet Parallel
AlexNet-based Image
Convert image pixels matrixes into Reduced Feature matrixes for all objects
227*2227*3
Multiplication of the (RL) of each dataset * each weight of AlexNet
∑ (Ri×inputs)+bias
Learned vectors
Extracted two types of vectors having all objects and their all features
Non-Learned with reduced feature vectors of all objects
Learned reduced feature vectors of all objects

Figure.15: Explainable Example 1 (Layer 3): Deep Learning Layer.


IR dataset
VIS dataset
RGB dataset
Modality Data 1
Modality Data 2
Modality Data 3
Learned reduced feature vectors of all objects
Learned reduced feature vectors of all objects
Learned reduced feature vectors of all objects
Value and weight data
Value and weight data
Value and weight data
Fusion level 2:Feature-based Fusions
Adaptive Fusion for object classification
Layer (4)
CNN-based Dempster-Shafer Fusion & object Classifier
CNN Dempster-Shafer and classified objects with relevant features
Probability assignment
Convolved feature
Pooled feature
Pooled feature for all objects
Learned reduced feature vectors of all objects

Sensor	Probability assignment										
	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}	U
$m_{CS}(\cdot)$	0.00	0.00	0.00	0.02	0.01	0.00	0.05	0.03	0.70	0.13	0.06
$m_{RE}(\cdot)$	0.00	0.00	0.25	0.25	0.00	0.00	0.00	0.00	0.25	0.00	0.25

Figure.16: Explainable Example 1 (Layer 4): Adaptive Fusion Layer.


IR dataset
VIS dataset
RGB dataset
Modality Data 1
Modality Data 2
Modality Data 3
Non-Learned with reduced feature vectors of all objects
Non-Learned with reduced feature vectors of all objects
Non-Learned with reduced feature vectors of all objects
Value and weight data
Value and weight data
Value and weight data
Fusion level 2:Feature-based Fusions
Adaptive Fusion for object classification
Layer (4)
CNN-based Concatenation Fusion & object classifier
CNN Concentration and classified objects with relevant features
Probability assignment
Concatenate
Concatenated features for all objects
Fused Concatenated Vector

Sensor	Probability assignment										
	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}	U
$m_{CS}(\cdot)$	0.00	0.00	0.00	0.02	0.01	0.00	0.05	0.03	0.70	0.13	0.06
$m_{RE}(\cdot)$	0.00	0.00	0.25	0.25	0.00	0.00	0.00	0.00	0.25	0.00	0.25

Figure.17: Explainable Example 1 (Layer 4): Adaptive Fusion Layer.

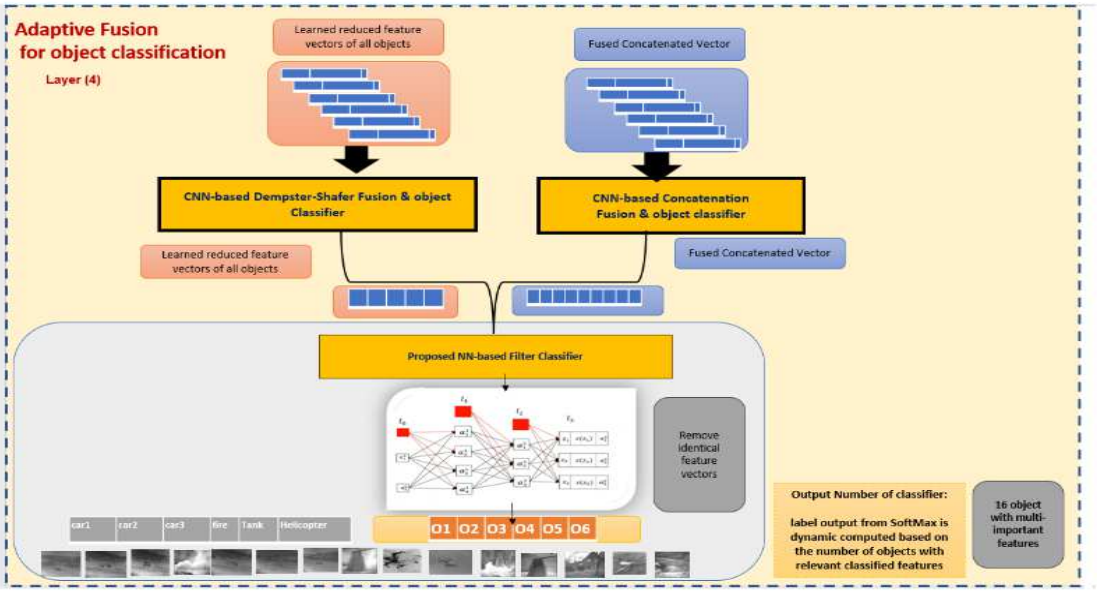
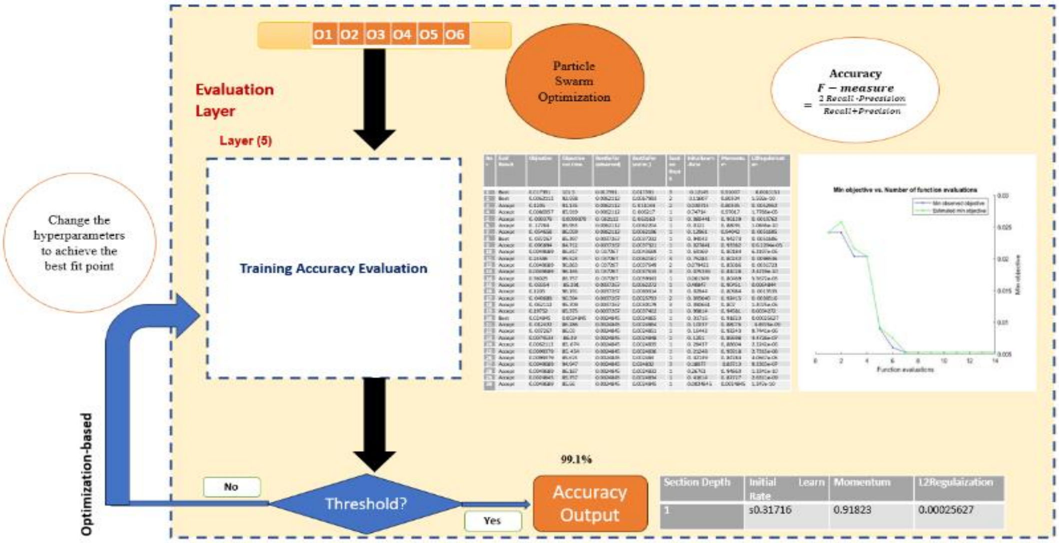

Adaptive Fusion
for object classification
Layer (4)
Learned reduced feature vectors of all objects
Fused Concatenated Vector
CNN-based Dempster-Shafer Fusion & object Classifier
CNN-based Concatenation Fusion & object classifier
Learned reduced feature vectors of all objects
Fused Concatenated Vector
Proposed NN-based Filter Classifier
Remove identical feature vectors
car1
car2
car3
fire
Tank
Helicopter
O1 O2 O3 O4 O5 O6
Output Number of classifier:
label output from SoftMax is dynamic computed based on the number of objects with relevant classified features
16 object with multi-important features

Figure.18: Explainable Example 1 (Layer 4): Adaptive Fusion Layer.


O1 O2 O3 O4 O5 O6
Particle Swarm Optimization
Accuracy
$F - measure = \frac{2\,Recall \cdot Precision}{Recall + Precision}$
Evaluation Layer
Layer (5)
Change the hyperparameters to achieve the best fit point
Training Accuracy Evaluation
Min objective vs. Number of function evaluations
Min observed objective
Estimated min objective
Function evaluations
Min objective
Optimization-based
No
Threshold?
Yes
99.1%
Accuracy Output

Section Depth	Initial Learn Rate	Momentum	L2Regulaization
1	s0.31716	0.91823	0.00025627

Figure.19: Explainable Example 1 (Layer 5): Evaluation Layer.

3.6.2 Explainable Example 2: (Different Modality Types: Text-audio)

This second different modality includes the five layers that are shown in Figure.20, Figure.21, Figure.22, Figure.23, Figure.24, Figure.25, Figure.26, Figure.27, and Figure.28. Figure 20 presents Explainable Example 2 (Layer 1): Software-Defined Fusion Layer. Figure 21 shows Explainable Example2 (Layer 2): Preprocessing Layer. Figure.22 shows Explainable Example 2 (Layer 3): Deep Learning Layer (part1). Figure.23 presents Explainable Example 2 (Layer 3): Deep Learning Layer (part2). Figure.24 presents Explainable Example 2 (Layer 3): Deep Learning Layer. Figure.25 shows Explainable Example 2 (Layer4): Adaptive Fusion Layer. Figure.26 presents Explainable Example 2 (Layer 4): Adaptive Fusion Accuracy. Figure.27 shows Explainable Example 2 (Layer4): Adaptive Fusion Accuracy. And Figure.28 shows Explainable Example 2 (Layer5): Evaluation Accuracy.

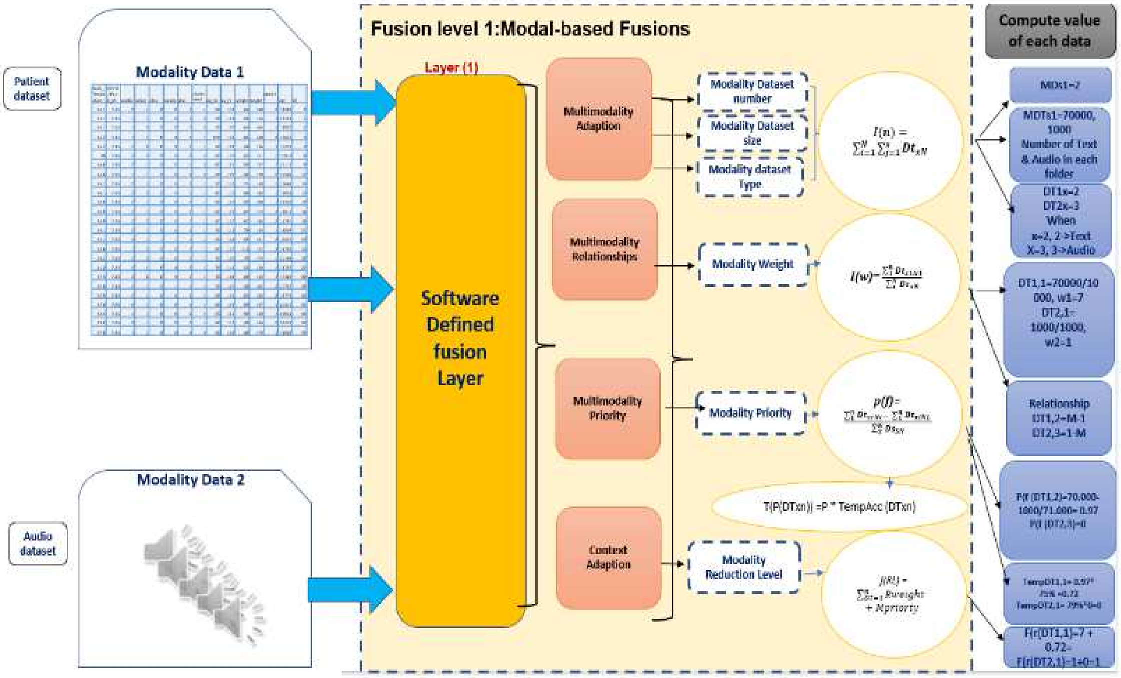

Patient dataset
Modality Data 1
Modality Data 2
Audio dataset
Fusion level 1:Modal-based Fusions
Layer (1)
Software Defined fusion Layer
Multimodality Adaption
Multimodality Relationships
Multimodality Priority
Context Adaption
Modality Dataset number
Modality Dataset size
Modality dataset Type
Modality Weight
Modality Priority
Modality Reduction Level
I(n) = ∑∑ Dt_xN
I(w)= ∑Dt / ∑Dt_xN
p(f)=
T(P(DTxn)) =P * TempAcc (DTxn)
f(R) = ∑ Rweight + Mpriorty
Compute value of each data
MDx1=2
MDTs1=70000, 1000 Number of Text & Audio in each folder
DT1x=2 DT2x=3 When x=2, 2->Text X=3, 3->Audio
DT1,1=70000/10 000, w1=7 DT2,1= 1000/1000, w2=1
Relationship DT1,2=M-1 DT2,3=1-M
P(f (DT1,2)=70.000-1000/71.000=0.97 P(f (DT2,3)=0
TempDT1,1= 0.97* 75% =0.72 TempDT2,1= 75%*0=0
F(r(DT1,1)=7 + 0.72= F(r(DT2,1)=1+0=1

Figure.20: Explainable Example 2 (Layer 1): Software-Defined Fusion Layer.

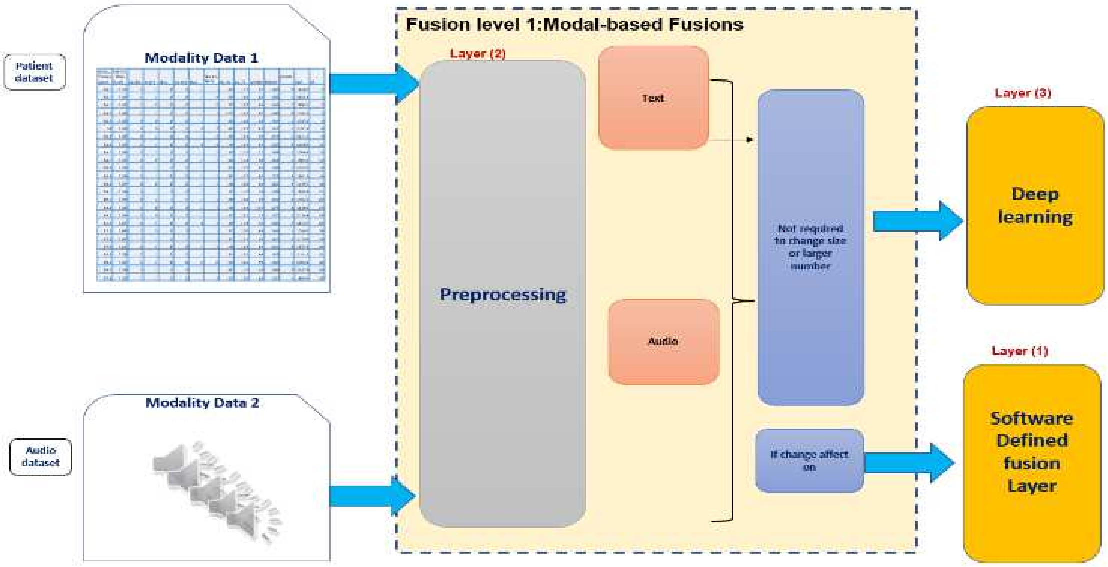


Figure.21: Explainable Example2 (Layer 2): Preprocessing Layer.

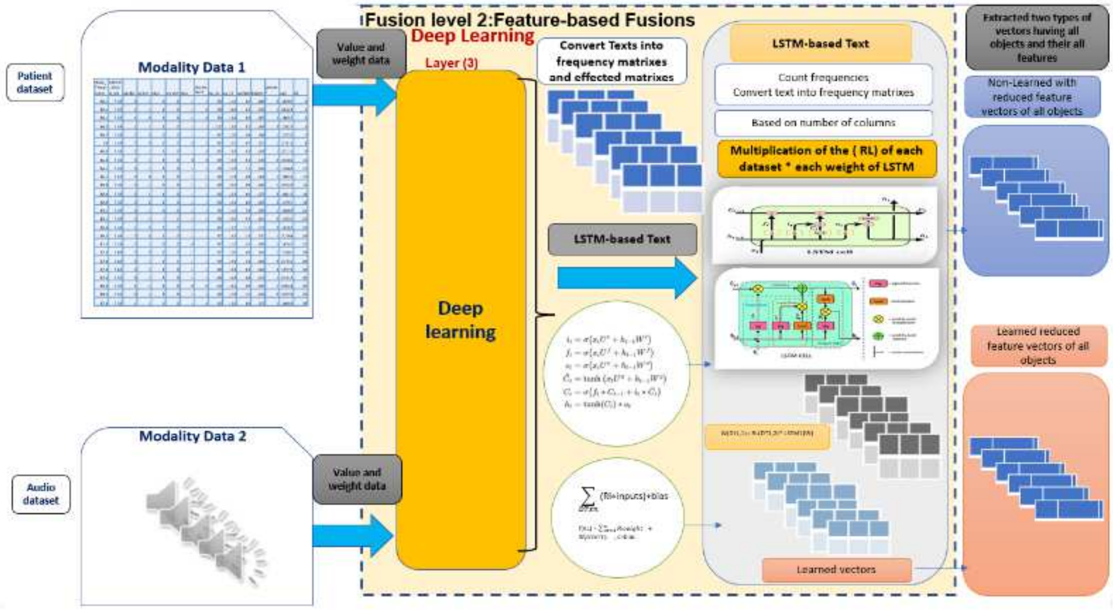

Patient dataset
Modality Data 1
Value and weight data
Fusion level 2:Feature-based Fusions
Deep Learning
Layer (3)
Deep learning
Convert Texts into frequency matrixes and effected matrixes
LSTM-based Text
LSTM-based Text
Count frequencies
Convert text into frequency matrixes
Based on number of columns
Multiplication of the (RL) of each dataset * each weight of LSTM
(fil+inputs)+bias
Learned vectors
Extracted two types of vectors having all objects and their all features
Non-Learned with reduced feature vectors of all objects
Learned reduced feature vectors of all objects
Modality Data 2
Audio dataset
Value and weight data

Figure.22: Explainable Example 2 (Layer 3): Deep Learning Layer (part1).

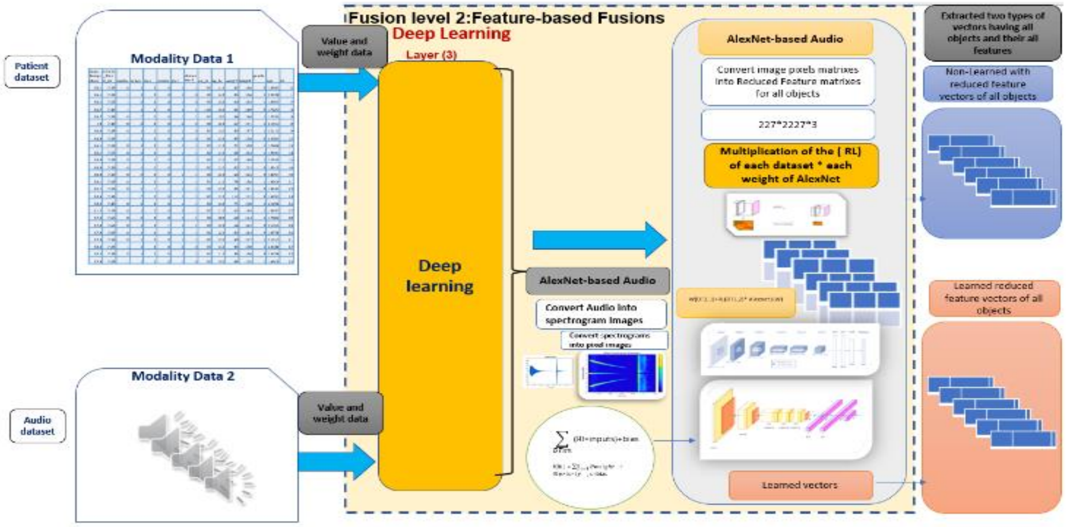

Patient dataset
Modality Data 1
Value and weight data
Fusion level 2:Feature-based Fusions
Deep Learning
Layer (3)
Deep learning
AlexNet-based Audio
Convert Audio into spectrogram Images
Convert spectrograms into pixel images
AlexNet-based Audio
Convert image pixels matrixes into Reduced Feature matrixes for all objects
227*2227*3
Multiplication of the (RL) of each dataset * each weight of AlexNet
(fil+inputs)+bias
Learned vectors
Extracted two types of vectors having all objects and their all features
Non-Learned with reduced feature vectors of all objects
Learned reduced feature vectors of all objects
Modality Data 2
Audio dataset
Value and weight data

Figure.23: Explainable Example 2 (Layer 3): Deep Learning Layer (part2).

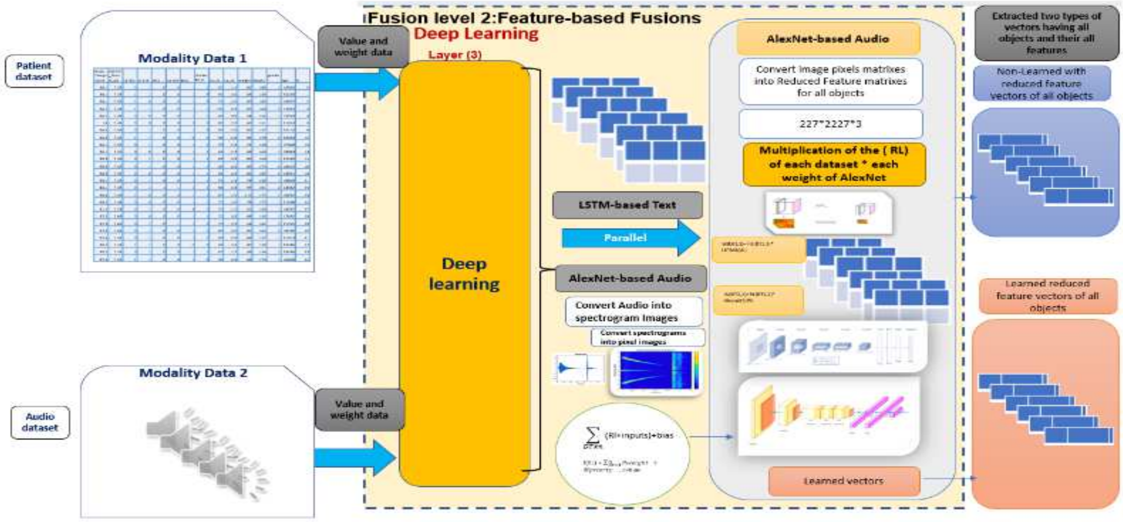
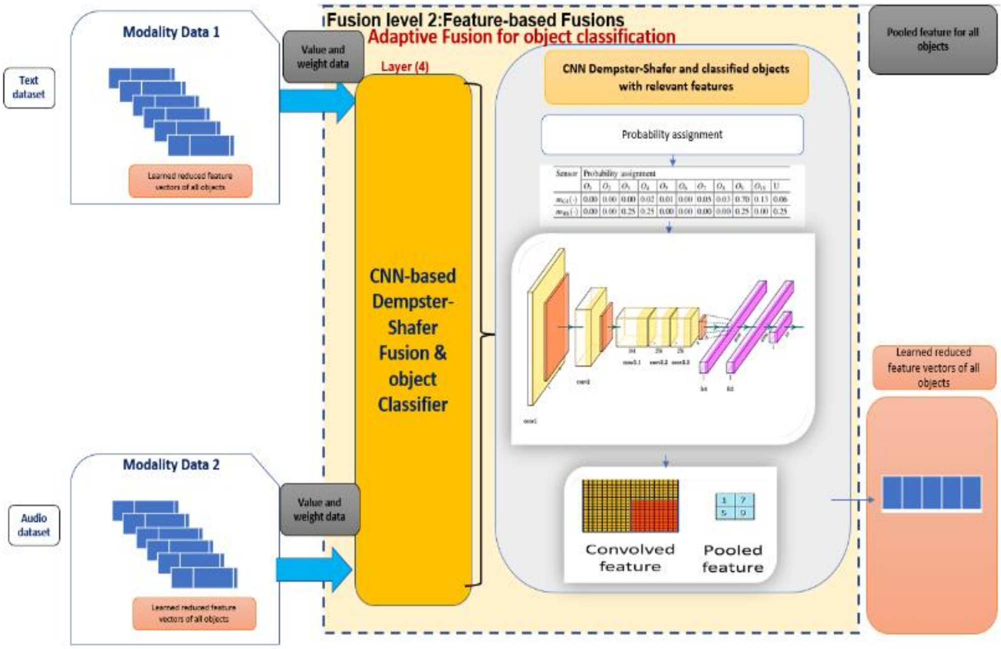
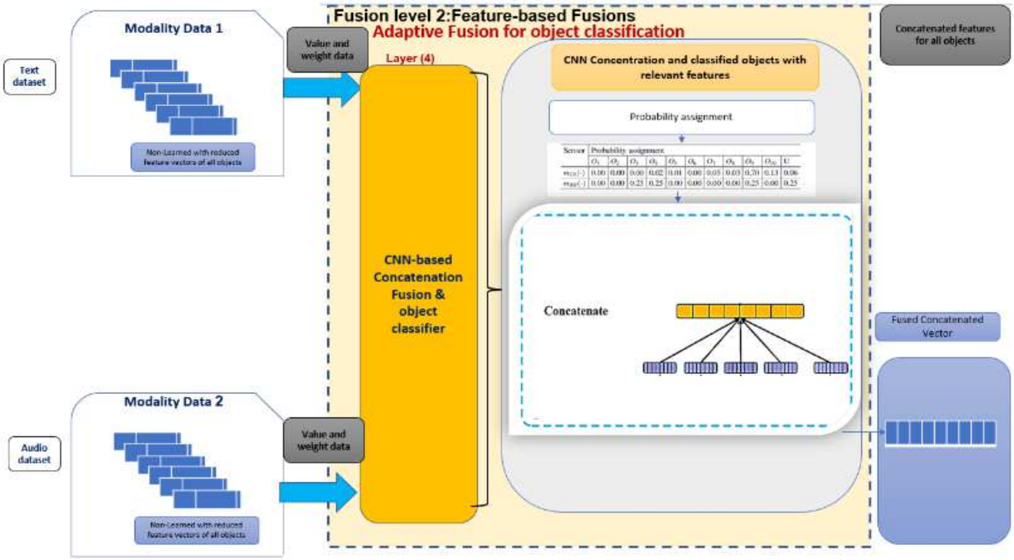

Fusion level 2:Feature-based Fusions
Deep Learning
Layer (3)
Patient dataset
Modality Data 1
Value and weight data
Deep learning
LSTM-based Text
Parallel
AlexNet-based Audio
Convert Audio into spectrogram Images
Convert spectrograms into pixel images
Modality Data 2
Audio dataset
Value and weight data
∑ (RI×inputs)+bias
AlexNet-based Audio
Convert Image pixels matrixes into Reduced Feature matrixes for all objects
227*2227*3
Multiplication of the (RL) of each dataset * each weight of AlexNet
Learned vectors
Extracted two types of vectors having all objects and their all features
Non-Learned with reduced feature vectors of all objects
Learned reduced feature vectors of all objects

Figure.24: Explainable Example 2 (Layer 3): Deep Learning Layer.


Fusion level 2:Feature-based Fusions
Adaptive Fusion for object classification
Layer (4)
Text dataset
Modality Data 1
Learned reduced feature vectors of all objects
Value and weight data
CNN-based Dempster-Shafer Fusion & object Classifier
CNN Dempster-Shafer and classified objects with relevant features
Probability assignment
conv1
conv2
conv3.1
conv3.2
conv3.3
Convolved feature
Pooled feature
1 7
5 0
Modality Data 2
Audio dataset
Learned reduced feature vectors of all objects
Value and weight data
Pooled feature for all objects
Learned reduced feature vectors of all objects

Sensor	Probability assignment										
	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}	U
$m_{CA}(\cdot)$	0.00	0.00	0.00	0.02	0.01	0.00	0.05	0.03	0.70	0.13	0.06
$m_{BS}(\cdot)$	0.00	0.00	0.25	0.25	0.00	0.00	0.00	0.00	0.25	0.00	0.25

Figure.25: Explainable Example 2 (Layer4): Adaptive Fusion Layer.


Fusion level 2:Feature-based Fusions
Adaptive Fusion for object classification
Modality Data 1
Text dataset
Non-Learned with reduced feature vectors of all objects
Value and weight data
Layer (4)
CNN-based Concatenation Fusion & object classifier
CNN Concentration and classified objects with relevant features
Probability assignment
Concatenate
Concatenated features for all objects
Fused Concatenated Vector
Modality Data 2
Audio dataset
Value and weight data
Non-Learned with reduced feature vectors of all objects

Sensor	Probability assignment										
	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}	U
$m_{CX}(\cdot)$	0.00	0.00	0.00	0.02	0.01	0.00	0.05	0.03	0.70	0.13	0.06
$m_{AE}(\cdot)$	0.00	0.00	0.25	0.25	0.00	0.00	0.00	0.00	0.25	0.00	0.25

Figure.26: Explainable Example 2 (Layer 4): Adaptive Fusion Accuracy.

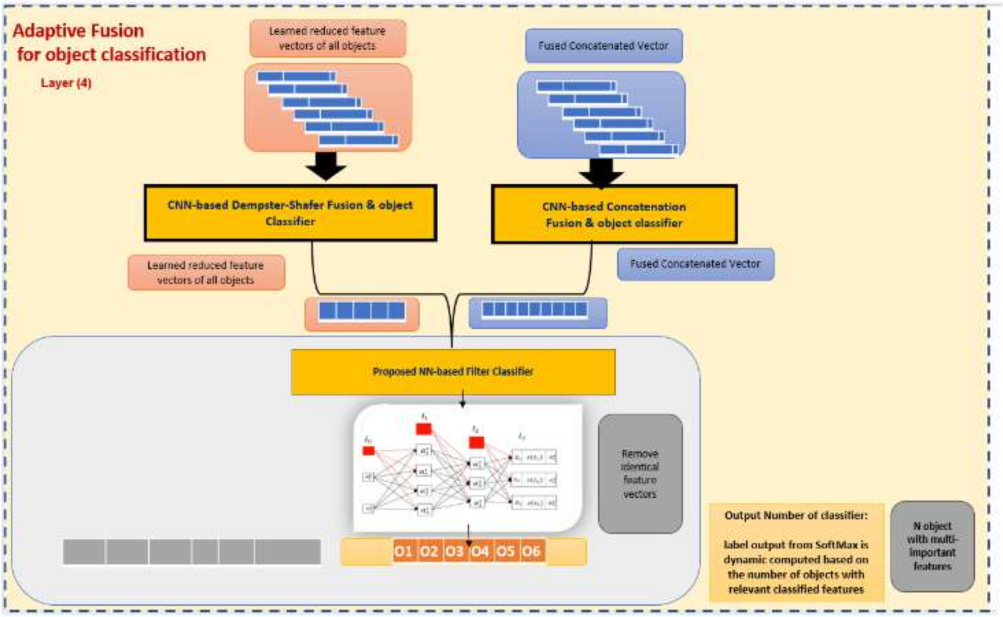
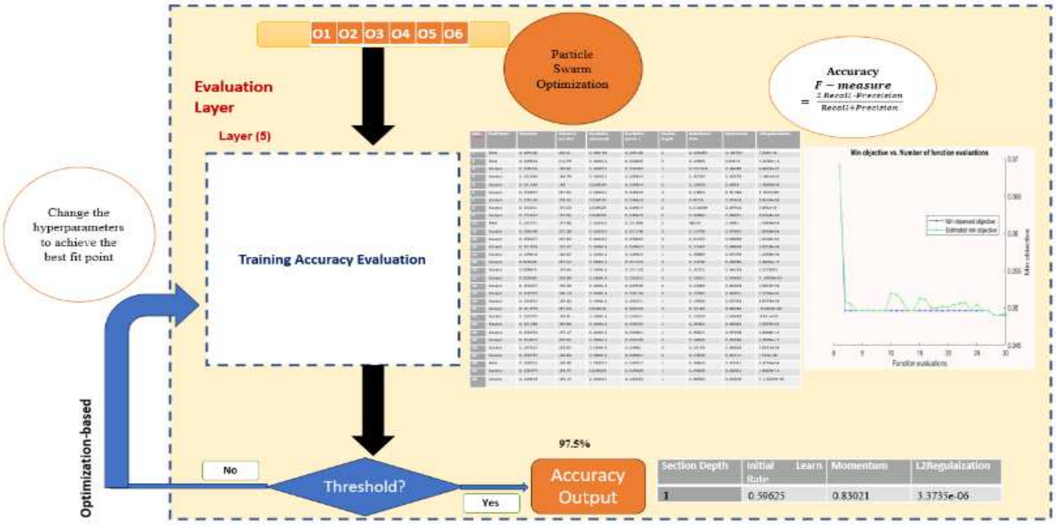

Adaptive Fusion
for object classification
Layer (4)
Learned reduced feature vectors of all objects
Fused Concatenated Vector
CNN-based Dempster-Shafer Fusion & object Classifier
CNN-based Concatenation Fusion & object classifier
Learned reduced feature vectors of all objects
Fused Concatenated Vector
Proposed NN-based Filter Classifier
Remove identical feature vectors
O1 O2 O3 O4 O5 O6
Output Number of classifier:
label output from SoftMax is dynamic computed based on the number of objects with relevant classified features
N object with multi-important features

Figure.27: Explainable Example 2 (Layer4): Adaptive Fusion Accuracy.


O1 O2 O3 O4 O5 O6
Particle Swarm Optimization
Accuracy
$F - measure = \frac{2\,Recall \cdot Precision}{Recall + Precision}$
Evaluation Layer
Layer (5)
Change the hyperparameters to achieve the best fit point
Training Accuracy Evaluation
Min objective vs. Number of function evaluations
Min observed objective
Estimated min objective
Function evaluations
Min objective
0.37
0.365
0.36
0.355
0.35
0.345
0
5
10
15
20
25
30
Optimization-based
No
Threshold?
Yes
97.5%
Accuracy Output

Section Depth	Initial Learn Rate	Momentum	L2Regulaization
1	0.59625	0.83021	3.3735e-06

Figure.28: Explainable Example 2 (Layer5): Evaluation Accuracy.

3.7 Implementation

Lastly, the proposed adaptive multimodal fusion framework is implemented using graphically user interface (GUI) to be easier and simpler. The proposed framework can be reusable for multiple information systems and mutual properties of multiple sources. We proof the proposed solution by implementing using MATLAB R2022 b that we construct the Adaptive smart Environment Multi-Modal System (ASEMMS) [64]. The goal of study Smart environment is not considerable specific context but it refers to interpreting the common big data characteristics of extracted data via multiple smart devices. The proposed solution has achievement of the fusion Objective that can fuse the interrelated of the complementary data for mange the data with unification objective. The implementation of proposed framework has five layers that is working parallel in two layers that are deep neural networks and fusion.

4. Experimental results

The experimental results explore the four experimental results based on the proposed adaptive smart environment multimodal framework that studies proposed MultiFusion models based on data perspective.

4.1 Measurements

The measurements rely on two dimensions, accuracy and optimization measurements.

a) Accuracy measurement

The accuracy evaluation measurement computes the classification accuracy of various classification models [65]. Precision-Recall is a useful measure of prediction success when the classes are imbalanced [66], [67],[68], as mentioned in Table.12. F1-measure is defined as the harmonic mean of precision and recall, as mentioned in equation (29), equation (30), and equation (31).

Table.12: The Accuracy Measurements.

	Conditions	True conditions	
Predicted condition	Predicted positive	TP	FP
	Predicted condition negative	FN	TN

The Equations for measuring the accuracy are defined as follows,

$$Precision = \frac{TP}{TP+FP} \qquad (29)$$

$$Recall = \frac{TP}{TP+FN} \qquad (30)$$

These quantities are also related to the F1 score, where

$$F - measure = \frac{2\ Recall \cdot Precsision}{Recall + Precision} \qquad (31)$$

Where: TP = True positive; FP = False positive; TN = True negative; FN = False negative. Precision (P) is defined as the number of true positives (TP) over the number of true positives plus the number of false positives (FP), as shown in equation (1). Recall (R) refers to the interpreted number of true positives (TP) over the number of true positives plus the number of false negatives (FN), as shown in equation (31).

B) Optimization Measurement

1. Bayesian optimization uses Bayes Theorem for optimizing expensive-to-evaluate functions that relies on black-box optimization [70].
2. Particle Swarm Optimization (PSO) depends on the mathematical formula to enhance the accuracy [71]. It is based on iteratively trying to enhance a candidate solution concerning a given quality evaluation. the PSO starts with input parameters, then generating first swarm, then measuring the suitable of all particles, determine best fitness of all particles, and find global best particle. lastly, after checking on swarm termination criteria that updates the position of particles and velocity of them as shown in equation (32).

$$v_i^{t+1} = v_i^t + c_1 U_1^t + (pb_i^t - p_i^t) + c_2 U_2^t + (gb^t - p_i^t) \qquad (32)$$

4.2 Datasets

Modality Fusion or Multi-modal Fusion refers to integration between multiple data types, or the same data type with various characteristics, Ambiguities and Inconsistencies, and Trivial Features. The proposed framework is generic for multimodality dataset in multiple contexts that adapts to fusion between multiple same or different modality types and number. The description of modality data types as shown in Table.13, Table.14, and Table .15. The description of modality datasets number in multiple contexts has a limitation of tracing from 1 to 16 modality input such as shown in Table.14 and Table.15. Table shows the sample of Multimodality datasets in Multicontext.

Table.13: Modality data type

Modality data type	Features
Text	Detected language, interpreted syntax, semantics, and syntax grammar.
Audio	Audio Speed, speech length, and kind of speech.
Image	Interpreted Dimension type, resolution of the image, and image modality type.
Video	The counted frames number and stream type whether real or offline.

Table.14: The Same Modality types with Diverse Modality Number study for diverse Smart environment systems

Modality number	**Multimodality**		
Same Modality types	**Bi-modal**	**Tri-modal**	**Multi-modal**
	Refers to two input modalities	Refers to three input modalities	Refers to more than three input modalities
	Image-Text fusion modality	Image-Image-Image fusion modality	Image- Image - Image - Image fusion modalities
	Text-Text fusion modality	Text-Text-Text fusion modality	Text-Text-Text-Text fusion modalities
	Audio-Audio fusion modality	Audio-Audio-Audio fusion modality	Audio-Audio-Audio-Audio fusion modality
	Video- Video fusion modality	Video-Video-Video fusion modality	Video-Video-Video-Video fusion modalities

Table.15: The Different Modality types with Diverse Modality Number Modality Number study for diverse Smart environment systems

Modality Number	**Multimodality**		
Different Modality Types	**Bi-modal**	**Tri-modal**	**Multi-modal**
	Refers to two input modalities	Refers to three input modalities	Refers to more than three input modalities

	Image-Text fusion modality	Image-Text-Audio fusion modality	Image-Text-Text-Video fusion modalities
	Image-Text fusion modality	Image-Text-Video fusion modality	Image-Image-Text-Video fusion modalities
	Image-Video fusion modality	Image-Text-Video fusion modality	Image-Image-Image-Text modalities
	Text- Audio fusion modality	Image-Video-Video fusion modality	Image-Text-Audio-Video fusion modalities
		Text-Text-Video fusion modality	Image-Text-Text-Video fusion modalities
		Image-Image-Text fusion modality	Image-Text-Text-Text Modalities
			Image-Audio-Audio-Text Modalities
			Image-Text-Text-Text Modalities
			Image-Image-Image-Video Modalities
			Image-Audio-video-Video Modalities
			Image-Image-Image-Text Modalities
			Image-Text-Audio-Text Modalities
			Image-Audio-Audio-Audio Modalities
			Image-Image-Video-Video Fusion Modalities
			Image-Text-Video-Video Fusion Modalities
			Image-Audio-Video-Video Fusion Modalities
			Image-Video-Video-Video Fusion Modalities

4. 3 Generic Experiments on Multimodality on Multicontext

This section presents a comparative accuracy analysis between a proposed adaptive fusion Model using Deep Learning and Dempster-Shafer fusion model and concatenation fusion model. The experiments to be generic and adapt with multimodality in multiple contexts, that interpret the data perspective of each data based on the target of complementary data whether interrelated data as patient's and meta data or the complementary fusion of the same objects in diverse datasets for example weapons datasets. This research classifies multimodality datasets based on interpreting the modality data types and number without known conditions and known context but all experimental dataset has the data criteria. The adaptivity of Multicontext is shown to be applicable in the diverse experimental datasets as smart military in three same modality input, smart health has two different modality input, Smart Diatrey health has three modality inputs. and Smart agriculture has four modality inputs as shown Table.16 and Table.17.

Table.16: The description of Multimodality datasets in Multicontext.

Context	Dataset	Modality dataset input	Modality dataset type	Modality dataset type
Smart military **[72], [73],[74]**	Multispectral Weapons objects	3 sources	Same (three sources)	3 images datasets
Smart health **[75], [76]**	Smart COVID-19 Health	2 sources	Different (two sources)	2 modality datasets (Text-Audio)
Smart health **[77]**	Smart Dietary Health	3 sources	Same type (three sources)	(Text- Text- Text)
Smart Agriculture [78]	Smart plant Diseases	16 sources	same Type (sixteen sources)	(16 Images sources)

Table.17: The detailed description of Multimodality datasets in Multicontext.

Context	Modality dataset input	Modality dataset Description	Modality dataset size	Modality dataset type	Modality dataset type
Smart military [72],[73], [74]	Dataset from Sensor of Insensitive spectrum (IR)	Weapons objects	10000 augmented images	Same	3 images datasets
	Dataset from Sensor of Visual insensitive spectrum (VIS)		10000 augmented images		
	Sensor RGB		10000 augmented images		
Smart COVID-19 health [75], [76]	COVID-19 patients Excel sheet	COVID-19 Patients	7000 records	Different	2 modality datasets (Text-Audio)
	Audios coughs datasets		1000 audios		
Smart Dietary health [77]	Smart Watch Excel sheet	Dietary patients	6265 records	Same	3 modalities (Text- Text- Text)
	Sensor Mobile Excel Sheet		3657 records		
	Sensor Mobile Images Folder for patients		2586 records		
Smart Agriculture [78]	Image dataset	Plant diseases	Training data 2.282.829.720	Same	16 modalities (16 Image sources)
	Image dataset				
	Image dataset		Validation data 571.061520		
	Image dataset folder				

Table.18: The sample of Multimodality datasets in Multicontext.

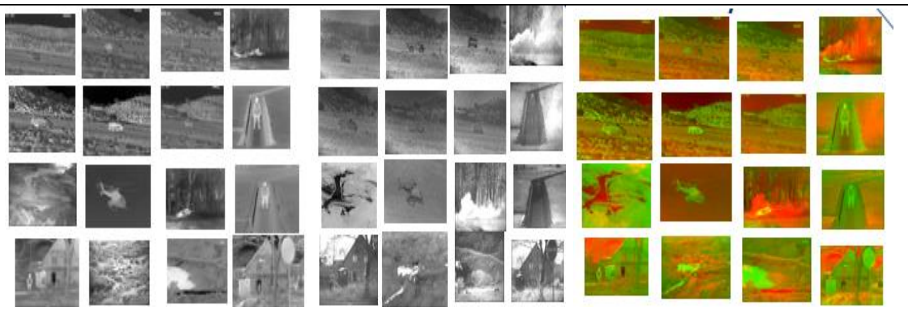
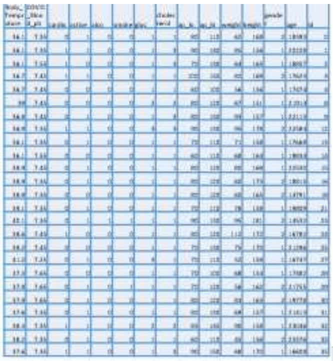
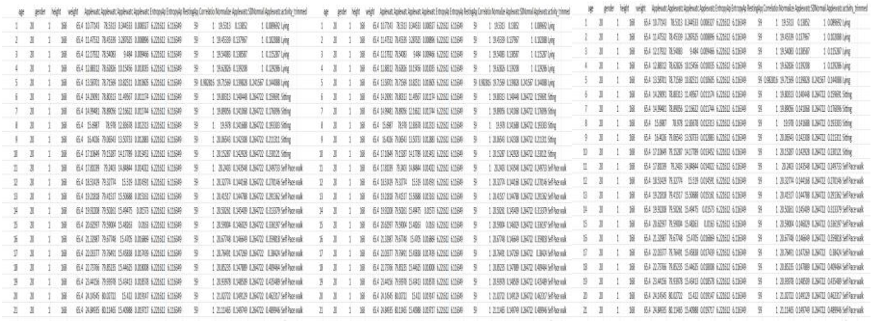
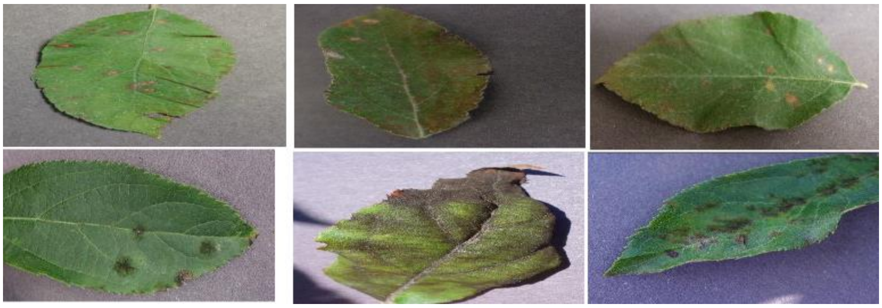
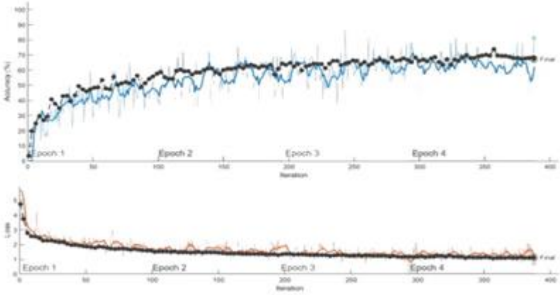
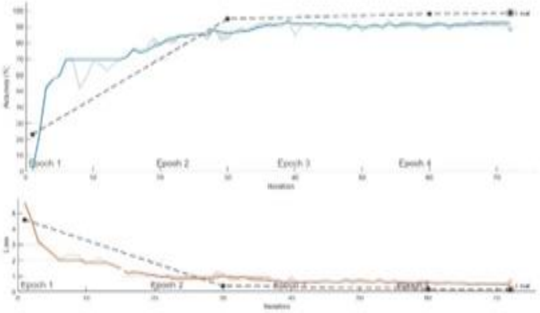
Context	Modality dataset input	Modality dataset Description	
Smart military [72], [73],[74]	Dataset from Sensor of Insensitive spectrum (IR)	Weapons objects	
	Dataset from Sensor of Visual insensitive spectrum (VIS)		
	Sensor RGB		
Smart COVID-19 health [75], [76]	COVID-19 patients Excel sheet	COVID-19 Patients	
	Audios coughs datasets		
Smart Dietary health [77]	Smart Watch Excel sheet	Dietary patients	
	Sensor Mobile Excel Sheet		
	Sensor Mobile Excel sheets Folder for patients		
Smart Agriculture [78]	Image dataset	Plant diseases	
	Image Dataset		
	Image dataset		
	Image dataset		
	Image dataset		
	Image Dataset		
	Image dataset		
	Image dataset		

	Image dataset		
	Image Dataset		
	Image dataset		
	Image dataset		
	Image dataset		
	Image Dataset		
	Image dataset		
	Image Dataset		

5. Results

The experimental results are classified into three analyses: (1) Proposed multimodal fusion Accuracy analysis with graphs. (2) the Best fit point of optimized accuracy with tracing 30 changes in hyperparameters in training for the proposed adaptive multimodal fusion framework accuracy results for four experimental results in diverse modality data types and number. And (3) Comparison between proposed fusion model and pervious two fusion models. The results are shown in Table.19 and Table.20. Table.19 explains the adaptive multimodal fusion framework accuracy results for four experimental results in diverse modality data types and number. Table.20 explores the adaptive multimodal fusion framework accuracy results for four experimental results in diverse modality data types and number.

Table.19: the adaptive multimodal fusion framework accuracy results for four experimental results in diverse modality data types and number.

Context	Accuracy
Smart military **[69], [70],[71]**	 The experimental accuracy result achieves to 98.8%
Smart health **[72], [73]**	 The experimental accuracy result achieves to 97.6%

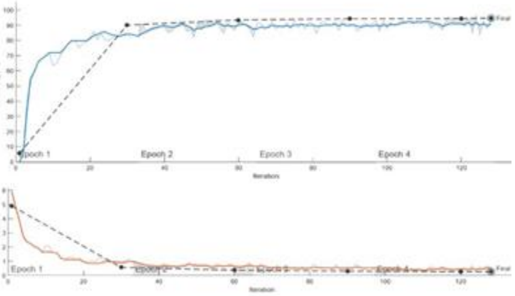
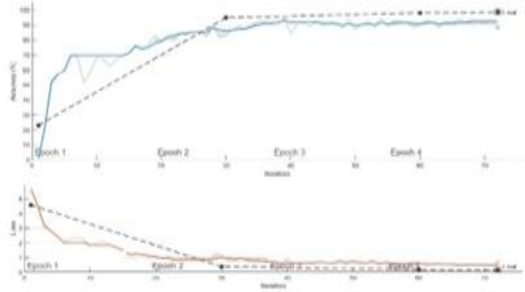
Smart health **[74]**	 The experimental accuracy result achieves to 95.9%
Smart Agriculture [75]	 The experimental accuracy result achieves to 98.5%

Table.21: a comparative analysis between the adaptive multimodal fusion framework accuracy results for four experimental results in diverse modality data types and number

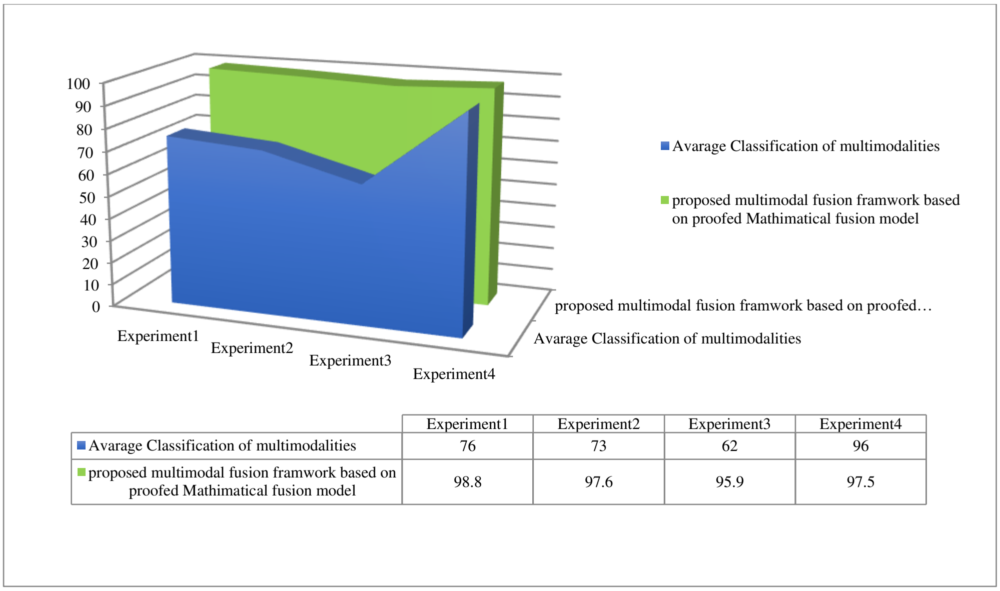
Context	Proposed Multimodal Fusion model of the adaptive framework
Smart military **[72], [73], [74]**	98.8%
Smart health **[75], [76]**	97.6%
Smart health **[77]**	95.9%
Smart Agriculture **[78]**	985%

6. Discussion

The proposed adaptive multimodal fusion framework can be applied on data or big data whether same or different via smart sensors or intelligent devices that can be applicable on the proposed data criteria in Preliminaries. This data can be applicable on one of suitable two types of data problems. Two types of suitable datasets can be applied with the proposed criteria, 1. Combine data of the same type from different sources to classify objects. For example (the thermal data of the weapons to get the best complete picture of each weapon object). 2. Combine multiple object data using different types or different characteristics to achieve the goal of combining the classification of objects that are often related to each other. For example (combining multiple patient X-rays recorded in an Excel sheet to create patient profiles and disease profiling based on the data). The discussion of the experimental results is categorized into three experimental analyses: (1) Comparison between accuracy analysis of average of multimodalities accuracies and proposed multimodal fusion Accuracy. (2) Proposed multimodal fusion Accuracy analysis with graphs. And (3) Comparison between proposed fusion model and pervious two fusion models. The results are shown the achieved average of proposed multimodal fusion framework accuracy is 97.45 with reduced feature level of multi-class of fused

MultiFusion learning model.

That interprets the proposed multimodal fusion framework is Better than contention fusion model by 28.5%. That interprets the proposed multimodal fusion framework is Better than Dempster-Shafer fusion model by 7.075%. The results are shown the achieved average of concatenation fusion model accuracy is 68.925 with a lot number of feature. The results are shown the achieved average of Dempster-Shafer fusion model accuracy is90.375 with limited features. Figure.29, Figure.30, and Figure.31 present the average of experiment results.



	Experiment1	Experiment2	Experiment3	Experiment4
🟦 Avarage Classification of multimodalities	76	73	62	96
🟩 proposed multimodal fusion framwork based on proofed Mathimatical fusion model	98.8	97.6	95.9	97.5

Figure.29: The behavior analysis of accuracy classification results and classification fusion results in many experiments for various modalities inputs.

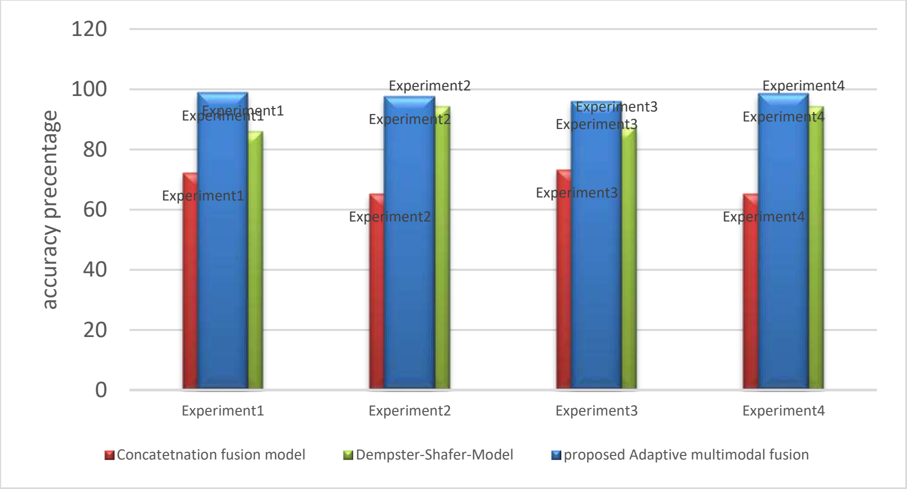


Figure.30: The behavior analysis of fusion techniques in many experiments for multi modalities inputs in Multicontext.

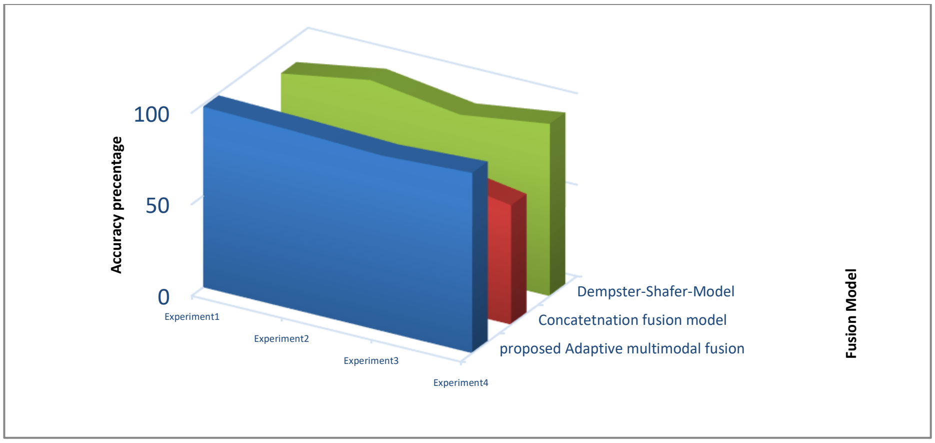


Figure.31: A comparative analysis of accuracy results between the proofed fusion model of the adaptive multimodal fusion framework and Concatenation fusion model and dempster-Shafer fusion model based on four experimental results in diverse modality data types and number.

7. Conclusion and future works

This paper presents the adaptive fusion framework that is a solution for the modality-context-based problem which is divided into two fusion problems, modality-based fusion and context-based fusion. The main challenge of modality-context-based is shown in the conflicting nature of data and complexity of fusion between sensory data. The modality-based fusion is interpreted into fusing multiple data sources with the same data type and fusing multiple diverse modality types via the same source in various smart systems. The context-aware can be described as interpreting the context that extracts relationships, features, conditions, and data modality types. The suitable datasets can be applied on the proposed criteria has two types, (1) Fusing similar types of data from various sources to object classification. For example (thermal datasets of weapons to achieve the best full vision of each weapons object's). (2) Fusing multi-target data via diverse types or different characteristics to achieve the unification target of object classification that are often interrelated to each other. For example (fusing multiple patients records Excel sheet with X-rays to make profiling of patient and profiling diseases from the data).

The proposed adaptive multimodal framework creates a MultiFusion learning model that can be adaptable and applicable to the same modality (Image, Text, Audio and Video) and to different multimodalities Multimodal fusion learning model has a vital role in fusing the complementary heterogeneous data to be reliable and robustness classification model in multiple contexts.

The proposed multimodal fusion framework is designed based on the proposed fusion taxonomy that It is considered the backbone of the proposed solution. Due to the proposed fusion taxonomy presents a full vision of any smart system with an unknown context based on data perspective. The proposed fusion taxonomy is built based on four dimensions, context type, reduction process type, data noise, and the flowing time of that data. This taxonomy is built on the basis of four criteria, which are modality data type, data reduction in relation to noisy or extraordinary data, and data temporal flow. The architecture of proposed adaptive multimodal framework has five layers, software-defined fusion layer, pre-processing layer, dynamic classification layer, adaptive fusion layer, and evaluation layer. It depends on inferring the deep neural networks and adaptive fusion layer for different modality types and modality input number. A software-defined fusion layer is as controller layer that can improve the classification accuracy by extracting relationships and infers weights and priority in the classification accuracy. A pre-processing layer is suitable dynamically for each modality type and working parallel. A dynamic classification layer creates four proposed deep learning models for four modality input types, image, text, audio, or video. An adaptive fusion improves the dempster-Shafer fusion theory by creating an adaptive fusion model between the dempster-Shafer and concatenation fusion. The adaptive fusion presents a fused full vision of modalities input and improves the accuracy classification results with getting bigger number of parameters and relationships. An Evaluation layer estimates the accuracy results and the optimization results in diverse the smart systems.

This paper applies multiple experimental results with multimodalities inputs in Multicontext to show the behaviors analyses of: (1) Proposed multimodal fusion Accuracy analysis with graphs. (2) the Best fit point of optimized accuracy with tracing 30 changes in hyperparameters in training for the proposed adaptive multimodal fusion framework accuracy results for four experimental results in diverse modality data types and number. And (3) Comparison between proposed fusion model and pervious two fusion models. (4) Comparison between accuracy analysis of average of multimodalities accuracies and proposed multimodal fusion Accuracy. (5) Proposed multimodal fusion Accuracy analysis with graphs. And (6) Comparison between proposed fusion model and pervious two fusion models.

There is growing interest in multidisciplinary research on multimodal synthesis technology to stimulate diversity of modal interpretation in different application contexts. The current literature review focuses on context-based systems in a specific known context and leaves a gap in incorporating multiple types of modal data in different contexts. Therefore, there seems to be a real need for an analytical review of recent developments in the field of data fusion. The real requirement for modality diversity across multiple contextual representation fields is due to the conflicting nature of data in multi-target sensors, which introduces other obstacles including ambiguous and inconsistent data. certainty, imbalance and redundancy in object classification. Additionally, there is a lack of frameworks that can analyze offline stream data to identify hidden relationships between different modal data types and different modal counts. Additionally, the lack of a multimodal fusion model capable of determining the extraction conditions of the extracted fusion data has resulted in low accuracy rates in object classification across modalities and systems, semantic system. This paper proposes a novel adaptive multimodal fusion framework for multimodal interpretation and contextual representation using evidence-enhanced deep learning Dempster-Shafer theory. The proposed framework is a solution to the open research challenge that is "context- and method-based fusion" to improve remote management, intelligent systems, and decision-making. The proposed framework can address the contradictory nature of data uncertainty, diversity of methods, factors, conditions and relationships for multimodal interpretation in multi-context systems.

The proposed multimodal fusion framework can be reusable across multiple information systems to improve decision making and control in different contextual representations. Furthermore, this study provides a comparative analysis between the current fusion model and previous multimodal data fusion models, explaining the differences between structural analysis, mathematical analysis of the model consolidation forms, their advantages and disadvantages. Furthermore, this study presents a comparative analysis between the proposed framework and three previous unified frameworks, exploring their concepts, advantages and problems, drivers, and current techniques. The results show that the average achieved accuracy of the proposed multimodal fusion framework is 98.45 at the multi-class feature level of the fused MultiFusion learning model. This renders the proposed multimodal fusion framework 28.5% better than the argument fusion model. This interprets the proposed multimodal fusion framework to outperform the Dempster-Shafer fusion model by 7.075%. The results show that the average chain fusion model accuracy achieved is 68.925 with many features. The results show that the average achieved accuracy of the Dempster-Shafer fusion model is 90.375 with limited functions.

Future work takes high attention to deeper analysis to particular fusion techniques. Briefly, new proposals to attempt different research directions, or simply inquisitiveness. There are several ideas that I would, the essential significant research direction is Fusion materials data science that expresses about the development of materials science in the industry has led to producing many materials data, which vary in data format and semantics and are extracted from multiple sources. The material data integration and fusion provide a unified framework for representation, processing, storage, and mining, which helps accomplish tasks such as material data clarification, material extraction, material fabrication parameter setting, and material knowledge extraction. And another research directions is Fusion from Big Data to Smart Data that can be discussed in the Smart data aims to filter noise data and produce valuable data, which can be effectively utilized by businesses and governments to plan, operate, monitor and control and make smart decisions. Although an unprecedented amount of data can be made available as advanced data merging technologies advance, the key is to discover how big data can become smart data and deliver intelligent information. Progressed huge information modeling and analytics are vital for finding implanted information frameworks and getting more brilliant information. Lastly, the future research direction is Multi-modal feature fusion by relational reasoning and attention for visual question answering that can be applied about Visual Address Replying (VQA) becomes to be a hot point in computer vision. A key arrangement to VQA exists in how to combine multi-modal highlights extricated from pictures and addresses. Previous motivation shows combining visual relationships and attention reaches more fine-grained feature fusion. Particularly, the importance of how to make a design to be an effective and efficient module in reasoning the complex relationship between visual objects.

4 References

1. Alberti, M.A., et al., Platforms for Smart Environments and Future Internet Design: A Survey, IEEE Access, Vol. 4, pp. 1-33, 2016.
2. Raun N.F., Smart environment using internet of things (IoTs) - a review, IEEE 7th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON), 2016.
3. The Statistics, Number of IoT devices 2015-2025 _ Statista, https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/

4. Zhi Y., et al., Deep transfer learning for military object recognition under small training set condition, Neural Computing and Applications, Springer, 2018.

5. Sundaravadivel; P., Kougianos; E., Mohanty; S.P., and Ganapathiraju, M.K., Everything You Wanted to Know about Smart Health Care: Evaluating the Different Technologies and Components of the Internet of Things for Better Health, IEEE Consumer Electronics Magazine, Vol.7 (1), pp:1-28, 2018.

6. Tunc , M.A., Gures, E., and Shayea, I.,A Survey on IoT Smart Healthcare: Emerging Technologies, Applications, Challenges, and Future Trends, arXiv:2109.02042v1 [cs.IT] ,2021

7. Nasr, M., Islam, M. Shehata, S., Karray, F., and Quıntana, Y., Smart Healthcare in the Age of AI: Recent Advances, Challenges, and Future Prospects, arXiv:2107.03924 [cs.CY], 2021.

8. liu, H., Deng, C., Fernandez-Caballero, A., and sun, F., Multimodal fusion for robotics,International Journal of Advanced Robotic Systems, Vol. 15(3):1, 2018

9. Hany F.A., and Robert J., and Gary W., Internet of Things: State-of-the-art, Challenges, Applications, and Open Issues, international Journal of Intelligent Computing Research, Vol. 9 (3), pp.928 - 938, 2018.

10. https://hevodata.com/learn/unstructured-data-definition/

11. https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/

12. Lahat, D., Adalı, T., and Jutten, C., Multimodal Data Fusion: An Overview of Methods, Challenges and Prospects. Proceedings of the IEEE, Institute of Electrical and Electronics Engineers, Multimodal Data Fusion, 103 (9), pp.1449-1477, 2015.

13. Wolter,D., and Kirsch.A., Smart Environments: What is it and Why Should We Care?. KI- Künstliche Intelligenz, 31 (3), pp.231-237, 2017.

14. Rashinkar, P., and Krushnasamy, V. S., An overview of data fusion techniques, International Conference on Innovative Mechanisms for Industry Applications (ICIMIA), 2017.

15. Baltrusaitis,T., Ahuja,C., and Morency,L-P., Multimodal Machine Learning: A Survey and Taxonomy,TRANSACTIONS OF PATTERN ANALYSIS AND MACHINE INTELLIGENCE, 2018, pp.1-20

16. Morency , L-P., Liang, P-P, and Zadeh, A., Tutorial on Multimodal Machine Learning,Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorial Abstracts, publisher:Association for Computational Linguistics, 2022, pp.33-38

17. Li, X., Eckert, M., Martine, J-F., and Rubio, G., Context Aware Middleware Architectures: Survey and Challenges, Sensors, Vol. 15(8), 2015 , pp.20570-20607

18. Khattak, A.M., Akba, N., Azam, M., Ali, T., Khan, A.M. Jeon, S., Hwang, M., and Lee, S., Context Representation and Fusion: Advancements and Opportunities, Sensors, Vol. 14, 2014,pp. 9628-9668;

19. Zaho, S., Gong, M., Fu, H., Tao, D., Adaptive Context-Aware Multi-Modal Network for Depth Completion, arXiv:2008.10833v1 [cs.CV], 2020

20. Furqan, A., Rashid , M., Iyad K., Nasser N.A., Data Fusion and IoT for Smart Ubiquitous Environments: A Survey, IEEE Access PP(99):1-1, 2017.

21. Atzori, L., et al., The Social Internet of Things (SIoT) – When social networks meet the Internet of Things: Concept, architecture and network characterization, Computer Networks, Vol.56 (16), 2012.

22. Žontar,R., Heričko,M., and Rozman, I., Taxonomy of context-aware systems, elektrotehniški vestnik Vol.79 (1-2), pp. 41-46, English Edition, 2012.

23. Baltrusaitis, T., Ahuja, A., Morency, L-P., Multimodal Machine Learning: A Survey and Taxonomy, IEEE Transactions on Pattern Analysis and Machine Intelligence (Volume: 41 (2), 2019,pp. 423-443

24. Panayides, A. S., Amini, A., Filipovic, N. D., Sharma, A., Tsaftaris, S. A., Young, A., Foran, D., Do, N., Golemati, S., Kurc, T., Huang, K., Nikita, K. S., Veasey, B. P., Zervakis, M., Saltz, J. H., & Pattichis, C. S. (2020). AI in Medical Imaging Informatics: Current Challenges and Future Directions. IEEE Journal of Biomedical and Health Informatics, 24(7), 1837– 1857. https://doi.org/10.1109/JBHI.2020.2991043

25. Barua, A., Ahmed, M.U., and Begum, S.,A Systematic Literature Review on Multimodal Machine Learning: Applications, Challenges, Gaps and Future Directions,IEEE Access, 2023

26. Lin Li, Chenkai Li, Xuanyu Lu, Hongmei Wang, Daming Zhou,Multi-focus image fusion with convolutional neural network based on Dempster-Shafer theory, Optik, Vol. (272), 2023.

27. Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L. Iuzzolino, Kazuhito Koishida, MMTM: Multimodal Transfer Module for CNN Fusion, Conference: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

28. Tang, X., Gu, X., Rao, L., and Lu, J.,A single fault detection method of gearbox based on random forest hybrid classifier and improved Dempster-Shafer information fusion, Computers and Electrical Engineering, Vol.92(2021), pp.1-18, 2021.

29. Mansoorizadeh, M., Charkari, N., Multimodal information fusion application to human emotion recognition from face and speech, Multimedia Tools and Applications, Vol. 49(2), pp:277-297, 2010.

30. Freitas, L.O., Henriques, P.R., and Novais, P., Context-Awareness and Uncertainty: Current Scenario and Challenges for the Future,International Symposium on Ambient Intelligence, 2018, pp.174-181

31. Khattak, A.M., Akba, N., Azam, M., Ali, T., Khan, A.M. Jeon, S., Hwang, M., and Lee, S., Context Representation and Fusion: Advancements and Opportunities, Sensors, Vol. 14, 2014,pp. 9628-9668;

32. Jenkins, M.P.; Gross, G.; Bisantz, A.M.; Nagi, R. Towards context-aware hard/soft information fusion: Incorporating situationally qualified human observations into a fusion process for intelligence analysis. In Proceedings of the 2011 IEEE First International Multi-Disciplinary Conference on Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), Miami Beach, FL, USA, 22–24 February 2011; pp. 74–81

33. Arwin D., Smart Military Society: Defining the characteristics to score the “Smart” of the military services, International Conference on ICT for Smart Society, 2013.

34. Goretarane, V., and Raskar, S., IoT Practices in Military Applications, Proceedings of the Third International Conference on Trends in Electronics and Informatics (ICOEI 2019), 2019.

35. Guo, Z., ET AL., A Feature Fusion Based Forecasting Model for Financial Time Series, Plos One, 2014. Brabandere, A.D., et al., Automating Feature Construction for Multi-View Time Series Data, Book: ECMLPKDD Workshop on Automating Data Science, pp.1-19, 2019.

36. Antonino Galletta, Lorenzo Carnevale, Alessia Bramanti, Maria Fazio,An innovative methodology for Big Data Visualization for telemedicine, IEEE Transactions on Industrial Informatics, 2018.

37. Diao, C., Wang, B., and Cai, N., A novel data fusion algorithm for multivariate time series, Chinese Control And Decision Conference (CCDC), 2018.
Xu, S., Chen, Y., Ma, C., and Yue, X.,Deep evidential fusion network for medical image classification, International Journal of Approximate Reasoning, Vol. 150, 2022, pp:188-198

38. Tang, X., Gu, X., Rao, L., and Lu, J.,A single fault detection method of gearbox based on random forest hybrid classifier and improved Dempster-Shafer information fusion, Computers and Electrical Engineering, Vol.92, 107101, pp.1-18, 2021

39. Che, C., Wang, H., Ni, X., and Lin, R., Hybrid multimodal fusion with deep learning for rolling bearing fault diagnosis, measurement, Vol.173 (7), 2020.

40. Raffaele G., et al., Multi-Sensor Fusion in Body Sensor Networks: State-of-the-art and research challenges, information fusion, 2016.

41. Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L. Iuzzolino, Kazuhito Koishida, MMTM: Multimodal Transfer Module for CNN Fusion, Conference: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020.

42. Zheng Tong, Philippe Xu, and Thierry Denoeus, An evidential classifier based on Dempster-Shafer theory and deep learning, arXiv: 2103.13549v1 [cs.AI], 2021

43. Abidin, R Z, Arshad, H., Shukri, S A A., Adaptive multimodal interaction in mobile augmented reality: A conceptual framework, Conference: the 2nd international conference on applied science and technology (ICAST’17), Vol. 1891(1), 2017

44. Heredia, J., Lopes-Silva, E., Cardinale, Y., Diaz-Amado, J., et al., Adaptive Multimodal Emotion Detection Architecture for Social Robots, IEEE Access, Vol.10, 2022.

45. Wagner, J., Fischer, V., Herman, M., and Behnke, S., Multispectral Pedestrian Detection using Deep Fusion Convolutional Neural Networks, Conference: 24th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN), 2016.

46. Canalle, G.K., Salgado, A.C., and Loscio, B.F., A survey on data fusion: what for? in what form? what is next?, Journal of Intelligent Information Systems , Vol. 57, pp. 25–50 , 2021.

47. Kampman, O., J. Barezi, E., Bertero, D., and Fung, P., Investigating Audio, Video, and Text Fusion Methods for End-to-End Automatic Personality Prediction, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2018.

48. Juan D. S. Ortega, Mohammed S., Eric C., Marco P., Patrick C., and Alessandro L. K., Multimodal Fusion with Deep Neural Networks for Audio-Video Emotion Recognition,Xiv:1907.03196v1 [cs.CV], 2020.

49. Wang, L., Luc, P., Recasens, A., Alayrac, J-B., and Oord, A.V.D., Multimodal Self-Supervised Learning of General Audio Representations, arXiv:2104.12807 [cs.SD], 2021.

50. Liu, J., Yuan, Z., and Wang, C., Towards good practices for multi-modal fusion in large-scale video classification, Computer Science, 2018.

51. Kelein, L., Mihaylova, L., and El Faouzi, N-E., Sensor and Data Fusion: Taxonomy, Challenges and Applications, In book: Handbook on Soft Computing for Video Surveillance, Edition: Taylor & Francis, Chapter: Sensor and Data Fusion: Taxonomy Challenges and applications, Publisher: Chapman & Hall/CRC, Editors: S. K. Pal, A. Petrosino and L. Maddalena, 2013.

52. Kuan, L., Yanen, L., Ning, X., Prem, N., learn to combine modalities in multimodal deep learning, arXiv:1805.11730v1 [stat.ML], 2018.

53. Al-Ateif, S., and Idri, A., Single-modality and joint fusion deep learning for diabetic retinopathy diagnosis, Scientific African, Vol. 17, 2022

54. Lu, Y., Zheng, W-L., Li, B., and Lu, B., combining eye movements and EEG to enhance emotion recognition, proceedings of the Twenty-fourth international on artificial intelligence (IJCAI), 2015.

55. LSTM

56. Juan D. S. Ortega, Mohammed S., Eric C., Marco P., Patrick C., and Alessandro L. K., Multimodal Fusion with Deep Neural Networks for Audio-Video Emotion Recognition,Xiv:1907.03196v1 [cs.CV], 2020.

57. Zhang, S., Zhang, S., Huang, T., and Gao, W., Multimodal Fusion with Deep Neural Networks for Audio-Video Emotion Recognition, ICMR'16: proceedings of the 2016 ACM on international conference on multimedia retrieval, pp.281-284, 2016.

58. D. WeimerAriandy , Benggolo, Y., Freitag, M., Context-aware Deep Convolutional Neural Networks for Industrial Inspection, Conference: Australasian Conference on Artificial Intelligence, At: Canberra, Australia, Volume: Deep Learning and its Applications in Vision and Robotics (Workshop) , 2015.

59. Leonardo, M.M., Carvalho, T., and Zucchi, R.A. Deep Feature-Based Classifiers for Fruit Fly Identification, Conference: 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 2018.

60. Che, C., Wang, H., Ni, X., and Lin, R., Hybrid multimodal fusion with deep learning for rolling bearing fault diagnosis, measurement, Vol.173 (7), 2020.

61. Ahmed, J., Muhammad, K., Won, S.i,K., Baik, S.W., and Rho, S., Dempster-Shafer Fusion based Gender Recognition for Speech Analysis Applications, IEEE, 2016.

62. Taheri, S., and Mammadov, M., Learning the naive Bayes classifier with optimization models, International Journal of Applied Mathematics and Computer Science, Vol. 23(4), 2013.

63. Zheng Tong, Philippe Xu, and Thierry Denoeus, An evidential classifier based on Dempster-Shafer theory and deep learning, arXiv: 2103.13549v1 [cs.AI], 2021

64. Doaa Mohey Eldin, Aboul Ella Hassanein , and Ehab E. Hassanien, ASEMMS: The Adaptive Smart Environment MultiModal System, Journal of System and Management Sciences Vol. 12 (2022) No. 2, pp. 1-20.

65. Gumawardama, A., and Shani, G., A Survey of Accuracy Evaluation Metrics of Recommendation Tasks, Journal of Machine Learning Research 10, pp. 2935-2962, 2009.

66. Taheri, S., and Mammadov, M., Learning the naive Bayes classifier with optimization models, International Journal of Applied Mathematics and Computer Science, Vol. 23(4), 2013.

67. Martinez-Ledesma, M., and Montoya, F.J., Performance evaluation of the particle swarm optimization algorithm to unambiguously estimate plasma parameters from incoherent scatter radar signals, Earth, planets, and space, Vol. 172, 2020.

68. Abadr, M., pourpanah, F., hussain, s., et al., A review of uncertainty quantification in deep learning: Techniques, applications and challenges, Information Fusion, Vol.76 , 2021, pp:243-297

69. Meng, T., Jing, X., Yan, Z., and Pedrycz, W., A survey on machine learning for data fusion, information fusion , Vol.57, 2020 , pp:115-129

70. Lau, B.p.l., Marakkalage, S.H., zhou, Y., hassan, N.U., Yuen, C., zhang, M., and Tan, x.,A survey of data fusion in smart city

71. Phoemphin, S., So-in, C., and Niyato, D., A Hybrid Model using Fuzzy Logic and an Extreme Learning Machine with Vector Particle Swarm Optimization for Wireless Sensor Network Localization, applied soft computing, Vol.65, 2018.
applications, vol.52, 2019, pp:357-374

72. TNO Image Fusion Dataset: https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029

73. Gun Dataset: https://www.kaggle.com/datasets/issaisasank/guns-object-detection

74. Flir Dataset FLIR Systems, Inc. FLIR ONE is Lightweight, Easy to Connect and Easy to Use, 2015

75. Cardiovascular Disease dataset: https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset

76. Respiratory Sound Database Dataset: https://www.kaggle.com/datasets/vbookshelf/respiratory-sound-database

77. Apple Watch and Fitbit data, https://www.kaggle.com/datasets/aleespinosa/apple-watch-and-fitbit-data

78. New Plant Diseases Dataset: ttps://www.kaggle.com/datasets/vipoooool/new-plant-diseases-dataset/data