Review Article

Advancements and Challenges in Deep Learning-Driven Marine Data Assimilation: A Comprehensive Review

Yunsheng Ma[1,2], Dapeng Zhang[1]* ⓘD, Yining Zhang[1], Guanyixuan Zhao[1], Yifan Xie[1], Haoyu Jiang[2]

[1] *Ship and Maritime College, Guangdong Ocean University, Zhanjiang 524088, China*
[2] *School of Electronics and Information Engineering, Guangdong Ocean University, Zhanjiang 524088, China*

Keywords	Abstract
	The acquisition and assimilation of high-quality data are fundamental for predictive model development across various domains. In the maritime realm, superior marine data fuels advancements in ship industry innovation, offshore clean energy initiatives, and marine engineering. Recent strides in employing deep learning methodologies have significantly improved data assimilation processes, raising the quality of derived datasets. This review meticulously examines deep learning-driven marine data assimilation, dissecting its challenges, identifying research gaps, and outlining future trajectories. This study employs Citespace's scientometric survey to comprehensively visualize and analyze the constituent elements within the literature, as well as to scrutinize the present state of research across pertinent fields, thereby providing an in-depth exploration and critical assessment of the scholarly landscape. Using bibliometric analysis, keyword exploration, and discipline classification, prevailing research patterns and emerging focal points are dissected. An insightful exploration into marine data nuances illuminates inherent challenges. Moreover, a comparative assessment of diverse algorithmic applications offers insights into their efficacy within this specialized domain. Culminating in a meticulous synthesis, this paper reveals pivotal developmental constraints in marine data assimilation, providing guidance for advancements across multifaceted dimensions in this field.

1. Introduction

Over the past few years, various technologies have been utilized for the research and development of marine resources as a result of technological advancements, including the use of artificial intelligence models to predict sea conditions, such as marine weather [1-3], and to study phenomena such as turbulence [4,5]; A marine system for the acquisition of marine energy and other resources [6,7]; Automated navigational systems for the navigation of ships [8-10], and so on. It is important to note, however, that these

technologies cannot be developed or optimized without a substantial amount of quality data. As sensors and other technologies provide incomplete, inaccurate, or noisy data, data assimilation techniques can help people obtain the highest quality data.

The concept of data assimilation refers to the process of integrating observed data with numerical model outputs [11]. Combining multiple data sources, including observations and model outputs, improves the estimation and prediction of the system state. In today's scientific research and practice, ocean data assimilation techniques play a crucial role. A

major component of our understanding of global climate change, marine resource development, and early warnings of marine disasters is the ocean, which is the largest ecosystem on earth. Although traditional ocean data assimilation methods offer some benefits in improving accuracy and credibility of ocean models [12], they have some limitations, including a lack of flexibility [13], difficulty in applying to complex systems with nonlinear dynamics [14], difficulty handling model-data mismatches [15], and vulnerability to errors and biases [16].

With the rapid development of computer science and technology, artificial intelligence technology has made great strides. Deep learning is one of the subfields of artificial neural networks that focuses on training these networks to perform tasks without explicit programming. Due to its superior ability to learn complex patterns and representations directly from data, it has gained immense attention and popularity. Adaptive model updating, outlier detection, and its ability to capture complex nonlinear relationships in data have made it a powerful tool for optimizing the limitations of traditional data assimilation techniques [17-20].

This article aims to advance the field. An analysis of the vast literature in this field is presented in the opening paragraphs of the article. Following the selection of literature, a review of the literature is conducted. In this paper, the characteristics of marine data are discussed, as well as the difficulties associated with their assimilation. More-over, the paper analyzes and compares the advantages and disadvantages of different deep learning approaches for data assimilation. In analyzing and discussing the review of this paper, we identify the challenges and gaps in the field as well as some potential directions and suggestions for future development. The paper can serve as a guide for the development of the field to some extent.

2. Data Access and Analysis

The articles screened in this subsection were analyzed bibliometrically. This analysis aimed to answer the following questions:

1) What are the most influential contributors (authors, countries)?

2) In this area of research, what are the recent and emerging frontiers of research?

3) What is the current application potential and scope of deep learning-based marine data assimilation techniques?

The schematic layout for the bibliometric survey of literature related to deep learning-based assimilation of marine data is shown in Figure 1.

The acceptance or rejection of the received manuscripts will be informed to the corresponding author and can be tracked by all authors through the journal web site. A paper which receives final or conditional acceptance, should be prepared regarding the requested corrections, and the revised manuscript should be resubmitted via the journal web site.
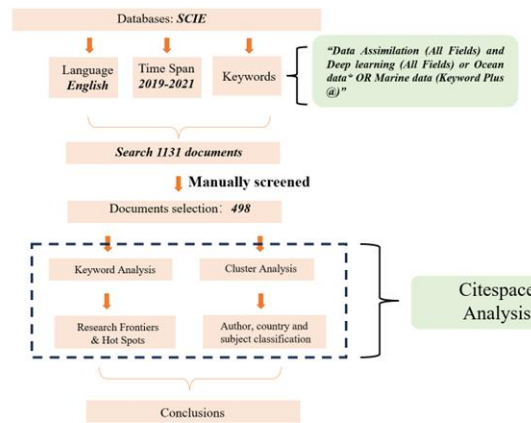

Figure 1. The flow diagram of the executed procedures for bibliometric review.

2.1. Data Access

The Web of Science Core Collection (WSCC) is a widely acknowledged and influential citation database that serves as a comprehensive repository of technical and scientific knowledge, playing a crucial role in facilitating effective data retrieval for scientometric analyses. In the pre-sent study, the Science Citation Index Expanded (SCIE) from the WOS Core Collection (WOSCC) database was deliberately selected. The dataset was acquired on May 2, 2023, from the online library of Guangdong Ocean University, China, utilizing an advanced search strategy: "Data As-similation (All Fields) and Deep learning (All Fields) or Ocean data* OR Marine data (Keyword Plus @)." This search method involved employing "Data Assimilation (All Fields) and Deep learning (All Fields)" for precise research field targeting, while the inclusion of "or Ocean data* OR Marine data (Keyword Plus @)" aimed to broaden the scope of the articles under consideration. Only articles published in English were included in the study, with proceedings, books, and re-views being excluded. Subsequently, the obtained results underwent manual review to eliminate irrelevant papers, resulting in the retrieval of 498 articles in plain text format, encompassing a comprehensive record of cited references for subsequent scientometric analysis.

The study employed CiteSpace, a sophisticated tool for scientific literature analysis and visualization. CiteSpace proved instrumental in assisting researchers in uncovering prevailing trends, research hotspots, and academic collaboration networks within their respective academic fields. The utilization of CiteSpace in this research underscores its significance as a valuable instrument for the systematic exploration of scientific literature, facilitating a nuanced understanding of the evolving landscape within academic domains.

2.2. Document Authors and Countries Analysis

Figure 2 shows that the most prolific contributors have published related articles. In terms of contribution, Lv is the most prominent contributor (14 articles), followed by Wang D. (6 articles), Zhang (5 articles), Li (5 articles), Wang B. (5 articles), Penny (5 articles), Hoteit (5 articles), and so forth. Also shown in Figure 3 is the extent to which different

countries contribute to the field. With 170 articles or 34.1% of the total number of articles, China is the largest contributor. This is followed by the United States with 165 articles or 33.1% of the total number of articles, France (57 articles, 11.4%), the United Kingdom (51, 10.2%), etc. Furthermore, the links in Figures 2 and 3 demonstrate author-to-author and country-to-country relationships. Clearly, there is a need to strengthen cooperation in this area.

Figure 2. Document author visual network in Machine learning based data assimilation research.
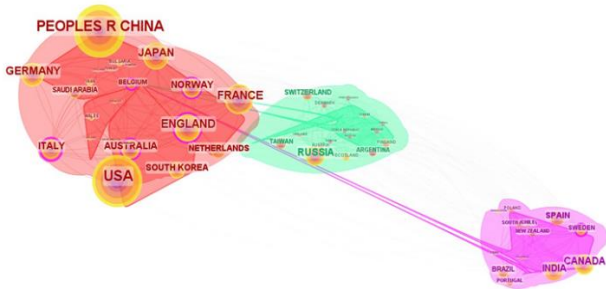

Figure 3. Document author visual network in Machine learning based data assimilation research.

2.3. Document Keywords and Categories Analysis

2.3.1. Keywords Analysis

Key research areas related to deep learning-based marine data assimilation can be identified through keyword co-occurrence analysis as they are distillation and core content of research articles. In order to visualize the knowledge graph of keyword co-occurrence analysis, it is essential to examine the change in frequency and centrality over time. There are 1555 links and 288 nodes in this network. In this dataset, each node represents a keyword, while the frequency of keyword occurrences shows the node, and the connections between keywords indicate a link. Figure 4 illustrates a deep learning keyword co-occurrence analysis network in the field of marine data assimilation. Keywords that are frequently

occurring are often close to one another, which is evidence of the main research in the field. In terms of frequency of occurrence, the main keywords are data assimilation, model, system, impact, and ensemble kalman filter. As a result of clustering, it can be determined that the current research hotspots in the field include data reconstruction, particle filtering, the North Atlantic, etc. Furthermore, a landscape view of Figure 5 illustrates the change in frequency of occurrence over time of the ten clusters of keywords. Using this figure, we can determine the research progress and the research hotspots over the last five years of research by analyzing and comparing the frequency of each period.
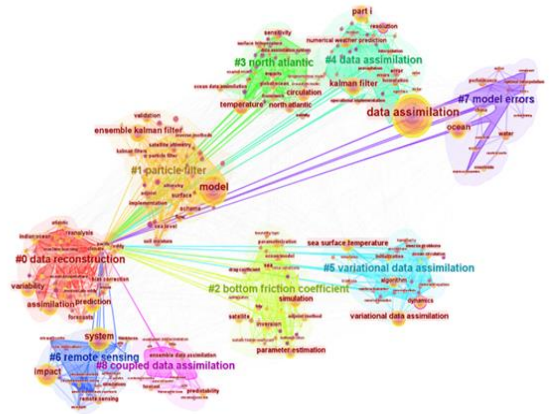

Figure 4. Machine Learning-Based Assimilation of Marine Data Related Research Keyword Co-occurrence Networks Knowledge Graphs.
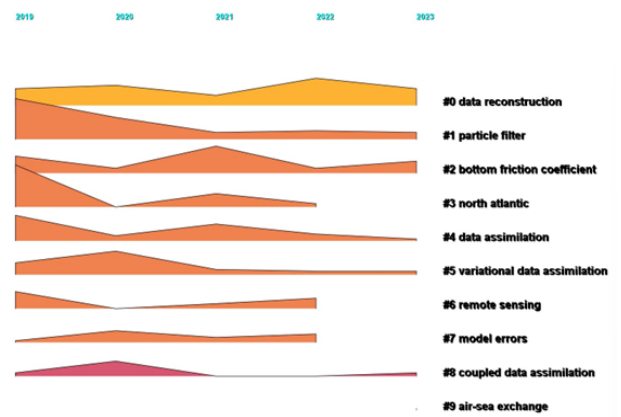
Figure 5. The landscape view of Machine Learning-Based Assimilation of Marine Data Related Research.

2.3.2. Scientific Categories Analysis

The disciplinary intersection of deep learning-based modeling of marine data assimilation with other research areas is well demonstrated by scientific categories. Here's a knowledge graph visualization for co-occurrence analysis of scientific categories. This network has 48 links and 106 nodes. In figure 6, you can see the co-occurrence analysis network for deep learning scientific categories. This field contributes to many fields like fluids and marine engineering, which in turn demonstrates the field's current state of research and development.

Figure 6. The co-occurrence analysis network for deep learning scientific categories.

3. Marine Data

In the marine environment, data are collected from a variety of sources, including oceans, seas, coastlines, and other saltwater bodies. There is a critical need for th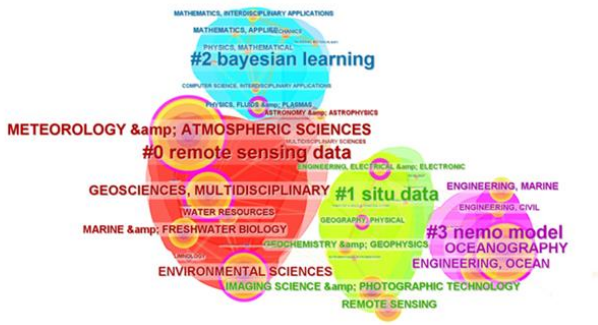ese data in order to understand and manage marine ecosystems, oceanographic processes, weather patterns, as well as a wide range of scientific, environmental, and commercial activities [21,22].

3.1. Main Types of Marine Data

Marine data can be divided into four main categories: oceanographic, bathymetric, meteorological and climatic, and biological and ecological data.

Oceanographic Data

A wide range of information is available about the physical, chemical, and biological properties of seawater in oceanographic data. There are several parameters that can be measured, including temperature, salinity, dissolved oxygen, nutrient concentrations, and currents. It is these data that are used by oceanographers to study ocean circulation, the distribution of marine organisms, and the interaction between the atmosphere and the oceans [23-25].

Bathymetric Data

Bathymetric data provide information about the topography of the seabed. Scientists and mariners use these data to understand underwater landforms, locate features such as seamounts or trenches, and plan activities such as submarine cable laying or resource exploration [26-28].

Meteorological and Climatic Data

The meteorological data include information about the weather conditions, such as temperature, humidity, wind speed, and atmospheric pressure over the oceans. Climate data refer to long-term trends and patterns in weather and atmospheric conditions. Climate change, storm forecasting, and maritime safety depend on these data [29,30].

Biological and Ecological Data

Data on biological and ecological organisms, ecosystems, and biodiversity are included in biological and ecological data. These data include information about species distribution, population dynamics, migration patterns, and the health of marine habitats. As a result of these data, conservation efforts, fisheries management, and research on the impact of human activities on marine life can be enhanced [31-34].

3.2. Methods for Obtaining Data

Currently, satellite observations, observations from ocean buoys and floats, observations from research vessels, and underwater instrumentation and sensors are the principal methods for collecting ocean data.

Satellite

Data such as sea surface temperature, ocean color, and sea level are collected by satellites equipped with remote sensing instruments. Data such as these are collected on a large scale and contribute to the understanding of global ocean dynamics [35-37].

Ocean Buoys and Floats

A variety of parameters, such as temperature, salinity, and currents, are collected in real-time by buoys and floats deployed throughout the oceans. By transmitting data back to researchers on land, these instruments contribute to continuous monitoring [38-41].

Research Vessels

A research vessel is a ship equipped with various instruments for the collection of marine data. Scientists can use them to examine oceanographic properties, geological features, and biological communities [42,43].

Underwater Instrumentation and Sensors

Data about the underwater environment is collected using autonomous underwater vehicles (AUVs) and remotely operated vehicles (ROVs). In addition to collecting data about ocean characteristics and marine life, these submersibles are capable of reaching depths that are difficult for humans to reach [44-46].

3.3. Characteristics of The Data

As a result of the complexity, scale, and dynamics of the marine environment. The characteristics of marine data are unique, including their spatial and temporal variabil-ity, multidisciplinary, volume, and complexity.

3.3.1 Spatial Variability

In the marine environment, there is a high degree of spatial variability, which means that data collected from different locations can differ greatly with respect to at-tributes such as temperature, salinity, and the distribution of marine organisms. It is necessary to use specialized techniques in the processing and analysis of such data in order to interpret and interpolate spatial differences [47,48].

3.3.2 Temporal variability

Ocean data are subject to temporal variability due to elements such as tides, seasons, and short-term events such as storms, which requires complex analysis to differentiate between natural fluctuations and significant trends. It is often necessary to analyze long-term datasets in order to identify meaningful patterns [49-50].

3.3.3 Multidisciplinary nature

Ocean data are derived from a variety of scientific disciplines. It is necessary to have an interdisciplinary approach when combining data from oceanography, meteorology, biology, and other disciplines to process and analyze the data effectively. As a result, integrating and interpreting data can be challenging [51].

3.3.4 Data volume and velocity

Ocean data are collected from a variety of sources, including satellites, buoys, re-search vessels, and underwater vehicles. Additionally, some data (such as meteorological and oceanographic data) are real-time in nature, requiring efficient processing and storage solutions to keep up with the constant influx of information [52].

3.3.4 Data quality and consistency

It is difficult to ensure the quality and consistency of ocean data because of factors such as sensor drift, calibration errors, and differences in data collection methods between platforms. In order to ensure the quality of data, data quality control and validation processes are essential, but they can be time-consuming [53].

4. Deep Learning Methods

Data assimilation based on deep learning is a relatively new approach that combines the principles of data assimilation with the power of deep learning algorithms. Combining observed data with numerical models improves the accuracy of state estimation and prediction in complex systems by utilizing neural networks.

Statistics are used in traditional data assimilation methods in order to combine model predictions with observed data while taking into account their uncertainties. Using deep learning-based data assimilation, neural networks are used to learn the relationship between model outputs, observed data, and the underlying dynamics of the system. Consequently, a flexible and nonlinear mapping between these components is possible, which may improve the performance of capturing complex system behavior [54-56].

4.1. Variational Autoencoders (VAEs)

A VAE is a generative model that learns possible representations of input data. Figure 7 illustrates the basic scheme of the variational autoencoder. These methods can be applied to data assimilation by learning compact, informative representations of a combination of observational and model data. Learned representations can be used for data assimilation and can reduce data dimensionality while maintaining important characteristics [57].
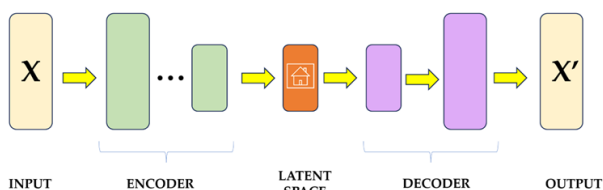

Figure 7. A variational autoencoder is based on the following basic scheme. Input X is provided to the model. In the encoder, it is compressed into the potential space. By sampling information from the potential space, the decoder produces X' as closely as possible to X.

In data assimilation, Ian presents a method for constructing analogues using variational autoencoders (VAEs). Through the use of VAE, the model state is mapped to the potential space by an encoder, and the vectors in the potential space are mapped back to the model state by a decoder. Based on the difference between the original and recon-structed model states, the parameters of the encoder and decoder are chosen to minimize the loss function [58].

In their study, Canchumuni et al. used a variational autoencoder to parameterize phase data from geological models and conditioned these models to observations using the ES-MDA method. However, VAE assumes that the potential space is continuous and generates new samples by interpolating within it. As a result of this interpolation, samples may be generated that cover uncorrelated features in the potential space, resulting in unreliable samples [59].

Yang et al. generated simulation ensembles using the Variational Autoencoder (VAE). Assimilation methods using VAEs are normally trained on the entire spatial domain, but they are problematic for complex, high-resolution models. In order to solve this problem, the authors split the state variables into multiple equal-sized patches and used VAE to encode and decode each patch in order to generate the simulation ensemble. As a result, discontinuities are reduced and errors are distributed uniformly over the entire spatial domain. As a consequence, generating the simulation ensemble using Variational Autoencoder (VAE) requires a considerable amount of training and computation, and training can take a considerable amount of time, especially for complex high-resolution models. Furthermore, choosing the right patch size is an important consideration; if the patch is too small, the model may not capture important features. The patch may also cause training difficulties and degrade the quality of the simulation ensemble if it is too large [60].

Cheng et al. used a variational autoencoder (VAE) to address the problem of chaotic latent spaces in autoencoders (AEs). The VAE augments the loss function with regularization terms and constrains the latent variables through Kullback Leibler Divergence (KLD) to ensure smoothness of the latent space geometry. This explicit latent space enhances the interpretability of the AE. The use of VAEs in autoencoders (AEs) may, however, result in loss of information due to the compression of the latent space. This may limit the accuracy of the model [61].

For the purpose of history matching, Zhang et al. used VAE to parameterize com-plex geological features. VAE is used to learn potential representations of phase dis-tributions and permeability distributions from geologic models. As a result of learning the latent representations, the VAE can obtain important information about the geologic features and produce new samples that are similar to the original ones [62].

4.2. Recurrent Neural Networks (RNNs)

Recurrent neural networks (RNNs) are neural network models capable of processing sequential data. As a result of its memory capability, it is able to utilize information from previous moments in order to influence the output of the present moment [63-65]. In order to integrate time-varying observations into models, they are able to capture temporal dependencies and patterns in the data. The use of these architectures is particularly advantageous when dealing with serial data, such as ocean currents, weather forecasts, or climate data.

A study by Penny et al. demonstrates the use of recurrent neural networks (RNN) in conjunction with data assimilation methods in numerical weather prediction (NWP) to replace computational forecasting models and enhance the accuracy of state estimation. As a pre-trained proxy model, the RNN can be initialized using DA methods to update the hidden reservoir state based on observations, allowing for repeated initialization of forecasts over a short period of time. By integrating the RNN with an ensemble Kalman filter and a 4D variational DA method, it is possible to accurately represent the system response to uncertainty under initial conditions and to assimilate sparse observations under uncertainty. Despite the absence of traditional numerical prediction models, RNN-DA methods can provide scalable and data-driven state estimation in NWPs and can be applied to higher dimensions through domain localization and parallelization. Nevertheless, the RNN-4dVar method is sensitive to sparse and noisy observation sets, which may affect its ability to estimate the state of the system. Furthermore, errors in the approximated background error covariance matrix and the RNN model equations used to derive the tangent linear model (TLM) may exacerbate the sensitivity to observation noise [66].

Using LSTM recurrent neural networks, Cheng et al. proposed a data-driven approach for improving the accuracy and efficiency of observation covariance specification in data assimilation for dynamical systems. As opposed to classical a posteriori adjustment method, this approach does not require knowledge or assumptions regarding the a priori error distribution. Observation covariance specification, assimilation accuracy, and computational efficiency are significantly improved by this method. Nonetheless, the method does not take into account the correlation pattern between the background and the observed error covariance, which may limit its ability to capture certain types of errors [67].

A Long Short-Term Memory (LSTM) network is a type of Recurrent Neural Network (RNN) designed to overcome the gradient vanishing problem inherent in traditional RNNs. A unit of LSTM can be seen in Figure 8. Using a reduced-order deep data assimilation (RODDA) model, Casas et al. integrated machine learning, dimensionality reduction techniques, and data assimilation. RODDA employs a long short-term memory (LSTM) network to model temporal dependencies and improve data assimilation accuracy. In addition, integrating machine learning, dimensionality reduction techniques, and data assimilation into RODDA models may add to the complexity of data assimilation [68].
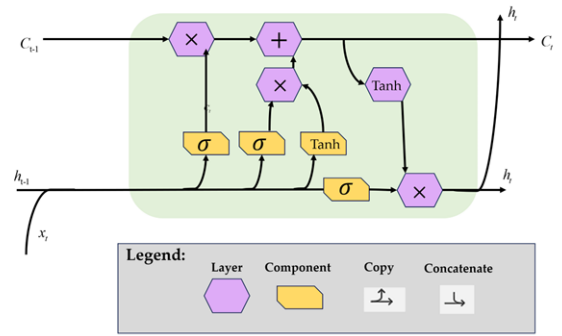

Figure 8. unit of LSTM

Deep Learning-Ensemble Kalman Filtering (DL-EnKF) leverages the capability of deep learning by embedding RNN models to improve the performance of EnKF in data assimilation. Through the learning of the dynamic features of the data, RNNs can be used to capture temporal dependencies in sequential data and improve the accuracy of data assimilation. The backpropagation algorithm may, however, encounter the problem of gradient vanishing or gradient explosion during the training process due to the cyclic structure of RNNs. As a result, the model may be difficult to convex or the training process may be unstable [69].

In Deep Data Assimilation (DDA), Data Assimilation (DA) and Machine Learning (ML) are combined to improve the accuracy of predictive models by reducing model errors. The DDA method uses a recurrent neural network that is trained using the state of the dynamical system and the results of the DA process in order to learn the assimilation process. The resulting model incorporates the features of the DA process and can be used for future predictions without reliance on DA. It is possible to apply the algorithms and numerical methods presented in this approach to other physical problems involving different equations and state variables. In spite of this, the DDA approach adds complexity to the prediction process since it involves training DNNs and integrating them with dynamic models. In order to implement and optimize the DDA algorithm, additional computational resources and expertise may be required [70].

4.3. Deep Neural Networks (DNN)

A deep neural network (DNN) is a deep learning model that consists of multiple layers of neural networks. As each layer extracts features from the input data, it gradually combines and abstracts them to achieve complex pattern recognition and prediction [71-73]. Its optimization in data assimilation is demonstrated by its ability to perform feature extraction, nonlinear modeling, large-scale data processing, and model generalization. Due to its powerful characterization and ability to learn features automatically, DNN is an important tool for data assimilation.

An optimal sensor placement strategy for data assimilation in turbulent flows based on deep neural networks. A feature-importance layer is incorporated into the DNN structure to determine the spatial sensitivity of the velocity to changes in the RANS model constants. In order to learn the relationship between the normalized velocity data and the model constants, the DNN model is trained

using the Adaptive Moment Estimation (Adam) algorithm. It is, however, important to consider the limitations and assumptions of the DNN model itself, as it may not be able to capture all of the complexity and complexities of turbulence. It is important to evaluate the generalization ability and robustness of the model under different flow conditions [74].

A data assimilation system that uses a deep neural network (DNN) as a likelihood function for detecting and classifying objects. In mathematics, a likelihood function is a function used to determine the probability of observing a set of data based on a set of input parameters. Probability functions based on DNN provide an alternative to likelihood functions based on the sum of squares of the differences between measured and simulated quantities. Similar to their application in image-based object detection and classification, DNN-based likelihood functions can be used to estimate uncertain variables in data assimilation. However, it is possible that the use of DNN-based likelihood functions may introduce additional uncertainty and bias since DNNs are trained on observed images and tested on simulated images, which may not fully reflect the complexity of real-world systems [75].

A new ensemble Kalman inversion method based on deep neural networks (DNNs) for modeling turbulence in separated flows at high Reynolds numbers. Data-driven eddy viscosity models are constructed using DNNs that are trained using experimentally accessible data such as velocity and force coefficients. To optimize the parameters of the model, gradient descent and back propagation algorithms are used [76].

Using deep neural networks (DNNs) to parameterize sub-grid scale processes in geophysical flows as part of a data assimilation method. Data assimilation techniques rely on deep neural networks to improve the prediction capabilities of multiscale sys-tems. Data-driven parametric models are built using deep neural networks to capture the mapping between resolved and unresolved variables in fluid dynamics. By using these models, it is possible to solve the closure problem in turbulence modeling and to improve the accuracy of prediction of complex physical systems. As a result of deep neural networks' tendency to overfit, they may perform well on training data but fail to generalize to new, unknown data [77].

A method of data assimilation that incorporates a deep neural network (DNN) for parameterizing sub-grid scale processes in geophysical flows. This approach enhances prediction capabilities and speeds up numerical simulations of these flows. Data assimilation techniques, in particular the Ensemble Kalman Filter, are used to calibrate the hybrid model during online deployment as an alternative to unresolved flow dynamics [78].

4.4. Convolutional Neural Networks (CNNs)

CNNs are well suited to tasks involving spatial data assimilation. In images or grid-based data, they excel at capturing spatial patterns and features. CNNs can be used to extract relevant spatial features from satellite imagery, oceanographic maps, or other geospatial data, improving the spatial accuracy of the model [79,80].

Specifically, Ruckstuhl et al. examine the application of Convolutional Neural Networks (CNNs) in Earth System Science, including data assimilation and improving weather and climate models. During data assimilation, Convolutional Neural Networks are used to preserve quality and positivity. In this study, CNNs were found to possess a high potential for improving the conservation of physical laws in data assimilation, and the feasibility of their use in complex numerical weather prediction systems was high-lighted. As a result, CNNs may not be able to generalize well to new and unseen data beyond their training distribution. In the real world, where data may exhibit different characteristics or uncertainties, this could limit the performance of CNNs [81].

CNN-PCA (Convolutional Neural Network Post-Processing with Principal Com-ponent Analysis) is a method for parameterizing complex 3D geographic models rele-vant to groundwater flow systems. It is a deep learning approach that combines the power of Convolutional Neural Networks (CNNs) and Principal Component Analysis (PCA) in order to represent complex geographic models using a small set of uncorrelated variables. CNNs are used to post-process PCA-parameterized models using the CNN-PCA procedure. For efficient data assimilation and history matching, this approach has been successfully applied to groundwater flow systems, including two-dimensional systems. The CNN-PCA approach, however, may have limitations in terms of its applicability to more complex systems and scalability to larger models [82].

In a study by Scott et al., ice/water observations can be efficiently retrieved from synthetic aperture radar (SAR) imagery of the Laurentian Great Lakes, Lake Erie, and Lake Ontario by means of a convolutional neural network (CNN). By learning features from the imagery, the CNN is able to reduce the number of coarse resolution training labels. In both dual-polarization and single-polarization retrievals, quality control measures based on the uncertainty of CNN outputs can effectively eliminate erroneous results. CNNs, however, have difficulty distinguishing between smooth, dark, and solidified ice cover regions. Data assimilation may be hampered by the misclassification of open-water observations in solidified ice sheets [83].

Predicting and estimating smallholder food production accurately is crucial to agricultural production. Consequently, smallholder farmers can estimate crop yields using an image-driven data assimilation framework. To estimate the probability distribution of rice crop states using images, convolutional neural networks are trained using labeled distribution learning [84].

An innovative network architecture combining a multilayer perceptron (MLP) and a convolutional neural network (CNN) enables the assimilation and inference of parameterized data. A CNN-SR model can be trained using only low-resolution samples instead of high-resolution labels because it is based on physics-based deep learning. Even if new entrance boundary conditions (BC) are introduced in the parameter space, the trained model can refine the spatial resolution of the flow field accurately [85].

4.5. Generative Adversarial Networks (GANs)

A GAN consists of a generator network and a discriminator network that compete with one another. By generating synthetic data samples that are consistent with both model predictions and observational data, GANs can be applied to data assimilation. In cases where observational data are scarce or noisy, this approach may be useful [86-88].

In order to handle data assimilation in non-Gaussian channelized aquifers, a new approach combines Deep Learning with Ensemble Smoother with Multiple Data Assimilation (ES-MDA). By assimilating hydraulic head and contaminant concentration data, the ES-MDA method is used to update the parameters of the potential space. With the help of Generative Adversarial Networks (GANs), generators can accurately re-produce the channelized structure with fewer parameters, lowering the uncertainty in hydraulic head and contaminant concentration predictions [89].

A multi-source information fusion generative adversarial network (MSIGAN) model is presented for parameterizing complex geologic features in history matching. To improve the accuracy of history matching, the model integrates a variety of information such as lithofacies distribution, micro seismic data, and well connectivity. The MSIGAN model combines the advantages of variational autoencoders (VAEs) and generative adversarial networks (GANs) in order to maintain geologic features during parameterization and history matching [90].

In epidemiology, Generative Adversarial Networks (GANs) can be used for spatiotemporal prediction (PredGaN algorithm) and data assimilation (dapredGan algorithm). As a result, GANs are set up within the framework of Reduced-order Models (NIROMs) in order to reduce the number of variables and make training easier. By using the proposed approach, it is possible to accurately predict the evolution of high-fidelity numerical simulations and to efficiently assimilate the observed data in order to de-termine the parameters of the model. In spite of this, the PredGan method is only able to interpolate prior data and does not attempt to extrapolate, which may limit its ability to make predictions beyond those observed [91].

SSIG-G is a generative adversarial network (GAN) model that derives daily sub-surface temperature fields from satellite remote sensing data. SSIG-G uses a convolutional neural network (CNN) as a generator to extract potential subsurface dynamical parameters and to learn the mapping from the input data to the real ST data. For the purpose of capturing complex hydrological features in satellite observations, it incorporates feature loss in the adversarial learning process. Under normal and extreme conditions, the SSIG-G model accurately represents physical oceanographic phenomena, providing high-quality inversion results. The generated data may, however, lack fine details and appear blurred, making high-resolution inversion difficult. It is a common challenge for traditional GANs that use only adversarial loss models [92]

4.6. Attention Mechanisms and Transformers

In assimilation, attention mechanisms, popularized by transformer models, can be used to weigh the importance of different data sources. As transformers can capture long-range dependencies and interactions between data points, they are particularly effective when handling sequential or spatiotemporal data [93,94].

3D-Geoformer, a transformer-based deep learning model, successfully predicts La Niña conditions for the second year of 2021 by representing the processes involved and using long time interval information as input to the predictor variables. Although 3D-Geoformer has demonstrated successful predictions, its limitations and uncertainties must still be taken into account when representing ENSO dynamics accurately [95].

ParaFormer is a training framework for hydrological parameter calibration. The framework consists of a transformer-based parameter learning model and an LSTM-based agent learning model. By using a self-attention mechanism, ParaFormer learns a global mapping from observed data to calibration parameters, capturing spatial correlations. Using the calibrated parameters as inputs, the agent model simulates ob-servable variables, such as soil moisture, overcoming the challenge of directly combin-ing complex hydrological models with deep learning technology [96].

A temperature prediction model based on Informer, a variant of Transformer, has been developed to improve the handling of time series data and to solve the long-term dependency problem in LSTM models. In time series forecasting, transformers, such as Informer, have emerged as potential solutions to the long-term dependence problem. In addition to enhancing the ability to predict long series, they also demonstrate excellent long-range alignment capabilities. Due to their ability to bypass the problem of long-term dependence, transformers are considered suitable for forecasting meteorological variables, including temperature. Due to the self-attention mechanism of transformers, they are able to predict each sequence element independently, making them more flexible when dealing with multiple inputs at once. The proposed model, however, does not include multivariate prediction, which limits its ability to predict all input variables simultaneously [97].

Data fusion model that integrates space station and radar data to predict precipitation, using a cross-attention mechanism to align and exchange feature information between the two modes. This model improves the accuracy and timeliness of the pre-diction and provides the flexibility to integrate other modal features. Based on four short-term rainfall datasets in southeastern China, it performed the best among the algorithms tested. The method, however, focuses solely on the integration of space station data and radar data without considering other modal features that may enhance prediction accuracy and timeliness [98].

4.7. Ensemble Learning with Deep Models

In data assimilation, ensembles of deep models can be used to capture uncertainty and variability. Several deep models could be included in the ensemble with varying architectures or initial conditions, and their outputs could be combined to provide a more robust assimilation result [99,100].

A method known as constructive simulation of ensemble optimal interpolation (CaneNOI). This method combines

generative models from machine learning with ensemble-optimal interpolation for data assimilation. In order to create ensemble members, generative models (e.g., generic and variational autoencoders) are trained on blocks of data. The ensemble members are then used in the data assimilation process. Although it is important to note that if the patch size is larger, it becomes more difficult to train accurate generative models, which may negatively affect the method's overall performance. Nevertheless, as patch size increases, data assimilation performance im-proves, resulting in a trade-off between accurate generative models and data assimilation [101].

To address simulator deficiencies, Luo proposed an ensemble-based learning framework to solve the supervised learning problem. As the data mismatch within each cluster is gradually reduced by the ensemble-based learning algorithm, the assimilation performance is improved. An ensemble-based method estimates a set of parameters, providing the benefits of ensemble-based methods, such as not requiring a complex and time-consuming concomitant system. Ensemble-based methods are also effective and derivative-free for estimating multiple sets of parameters, allowing uncertainty quantification [102].

He et al. combined the physics agnostic data-driven stochastic feature map approach as a predictive model for ensemble Kalman filter data assimilation. Ensemble Kalman filters combine multiple predictions (called ensemble members) to improve prediction accuracy. A machine learning model is learned sequentially by integrating incoming noisy observations, and the predictive model obtained exhibits a very high level of predictive ability. Additionally, the method can be used to generate reliable ensembles for probabilistic predictions [103-107].

5. Discussion

In the above section, recent advances in the area of optimizing deep learning for data assimilation techniques are reviewed. The above review indicates that deep learning techniques have made significant advances in optimizing data assimilation methods across a wide range of domains. The purpose of data assimilation is to combine observed data with model simulations in order to provide a more accurate and up-to-date estimate of the state of the system. Data assimilation can be enhanced significantly by deep learning. At this stage, the following advantages are available: im-proved state estimation; nonlinear system modeling; and enhanced data quality control.

5.1. Difficulties and limitations

There are still some limitations and difficulties in the field of deep learning, although it has shown great promise for optimizing data assimilation methods. The paper discusses in detail the difficulties identified in the above review in this subsection.

5.1.1. Data Requirements

The data requirements for data assimilation based on deep learning may pose specific limitations and challenges which have hindered the development of this field to some extent. As a result, they should be addressed with care.

- **Data Quantity**

There may be a lack of observational data in some data assimilation scenarios, which is critical for the calibration and validation of models. It can be challenging to obtain comprehensive real-world measurements for certain environmental or geospatial variables.

- **Data Consistency**

It's possible for data from different sources to have temporal mismatches, meaning observations were taken at different times than model predictions. It's hard to deal with such temporal discrepancies.

- **Data Diversity**

Diverse data sources and modalities are beneficial to some deep learning models. An insufficient variety of data types can prevent the model from fully capturing the complexity of the system (e.g., satellite imagery, ground-based measurements, remote sensing data).

- **Data Representativeness**

A model may not be able to assimilate all conditions and events based on observations. Under certain circumstances, biased sampling can result in poor model performance.

- **Historical Data**

It is possible that historical data may be limited or outdated in some cases. Large and current datasets are often crucial to the development of deep learning models. Model training and validation can be hindered by a lack of historical data.

5.1.2. Overfitting and Generalization

In machine learning and deep learning, overfitting and generalization are critical concepts. It is important to keep these factors in mind when applying deep learning techniques to tasks relating to data assimilation, even though they are not unique to data assimilation.

- **Data Scarcity**

An important challenge in data assimilation is obtaining a large and diverse dataset for model training, especially when dealing with spatiotemporal environmental data. A lack of data can result in overfitting, in which the model captures the noise in the training data rather than the underlying patterns.

- **Complex Models**

Models based on deep learning can contain millions of parameters, making them highly flexible and capable of fitting noisy data. As a result of this flexibility, there is an increased risk of overfitting, particularly when the model is unable to capture the true underlying relationships because its capacity exceeds what is necessary.

- **Hyperparameter Tuning**

In order to mitigate overfitting, deep learning models require careful tuning of several hyperparameters, including the number of layers, the learning rate, and the dropout rate. It can be time-consuming and resource-intensive to determine the appropriate set of hyperparameters.

- **Imbalanced Data**

It is possible that some variables or regions have more data than others in some data assimilation applications. An overfitting error occurs when the model overemphasizes well-sampled areas and underperforms in sparsely sampled areas.

- **Short Training Sequences**

In time series data assimilation, particularly in the case of phenomena that change rapidly, such as the weather, short training sequences can lead to overfitting. With limited historical data, the model may be unable to capture long-term dependencies.

- **Limited Model Generalization**

When trained on a specific dataset or domain, deep learning models may not generalize well to new conditions or regions. It is possible for models that are trained for a specific geographic region or time period to underperform when they are applied to different regions or time periods.

- **Environmental Variability**

The behavior of environmental and geospatial systems can be complex and variable. A model that does not generalize well may not capture the full range of variability in the system, leading to inaccurate results from the assimilation process.

5.1.3. Interpretability

There is a tendency for deep learning models to be viewed as black boxes, making it difficult to interpret the reasoning behind their predictions. Interpretability is essential for understanding model behavior and ensuring physical consistency in data assimilation.

- **Complex Model Architectures**

In deep learning models, especially deep neural networks, millions of parameters can be taken into account, as well as complex architectures with many layers. As a result of this complexity, it may be difficult to interpret how each parameter contributes to the decision-making process of the model.

- **Non-linearity**

The nature of deep learning models is inherently non-linear. Using these methods, you can capture intricate and non-linear patterns in data that are difficult to visualize and explain using traditional linear methods.

- **Black-Box Nature**

The internal workings and representations of many deep learning models are considered black boxes, which makes it difficult to understand their internal workings. It may be difficult to determine the relationship between inputs and outputs.

- **High-Dimensional Data**

When dealing with high-dimensional data, such as images or text, interpretation becomes more challenging because it may be difficult to determine which features or combinations of features the model is focusing on.

- **Interactions Between Features**

It is difficult to attribute a specific model decision to a specific feature or combination of features in a deep learning model due to the complex interactions between features.

- **Transferability of Interpretations**

Interpretations generated for one deep learning model may not necessarily be applicable to another model that uses a different architecture or dataset. Model-specific interpretations are possible.

- **Trade-off with Performance**

It is important to note that some techniques used to enhance interpretability, such as simplifying the model architecture or using interpretable surrogate models, may lead to a reduction in model performance.

5.1.4. Model Complexity

In operational settings, deep learning models can be difficult to implement, maintain, and fine-tune for specific assimilation tasks due to their complexity.

- **Large Training Datasets**

It is necessary to use large training datasets for complex models in order to prevent overfitting, which may not be available in all domains or may be extremely expensive.

- **Time-Intensive Training**

Developing complex models is time-consuming, limiting the ability to rapidly iterate and experiment with different model architectures.

- **Scalability**

It is possible that complex models may not scale well for deployment on edge de-vices or for use in real-time applications due to their computational demands, which may limit their usefulness.

5.1.5. Domain specificity

There is a possibility that deep learning models may not generalize well across different domains or environmental conditions. Their customization and fine-tuning are often domain-specific, which makes them less adaptable to a variety of assimilation scenarios.

- **Scalability**

Due to the fact that domain-specific models are designed to solve a specific problem, they may not perform well when they are applied to a broader range of sources or to larger datasets.

- **Lack of Interoperability**

The integration of domain-specific models into existing systems or workflows can be challenging, particularly when dealing with models from different domains.

- **Maintenance Challenges**

It is possible that domain-specific models may suffer from model decay over time as the domain evolves. Maintaining and updating these models can be a time-consuming and resource-intensive process.

- **Interdisciplinary Gaps**

It is often necessary to have expertise in both machine learning and the specific domain when developing domain-

specific models. It can be challenging to bridge the gap between these disciplines.

5.1.6. Assimilation of Uncertainty

The point estimates provided by deep learning models may not adequately capture uncertainty. In data assimilation, accurate representation and assimilation of uncertainty are crucial, and deep learning approaches can present challenges in this regard.

- **High Computational Demands**

In complex systems, modeling uncertainty can require computationally intensive strategies, such as Monte Carlo simulations or Bayesian methods, making it impractical for real-time applications or large datasets.

- **Model Error**

Errors in model assumptions: Assimilation techniques often rely on models that simplify underlying physical or biological processes. Assimilation can be affected by errors and uncertainty caused by these assumptions.

- **Non-Gaussian Distributions**

In real life, there are a number of uncertainty sources that do not follow Gaussian distributions, which are commonly assumed in assimilation techniques. Non-Gaussian uncertainty can be complex and computationally intensive to handle.

- **Parameter Uncertainty**

Estimating uncertainty in model parameters, such as coefficients or initial conditions, can be challenging, and may require additional data and calibration.

- **Sensitivity to Initial Conditions**

The butterfly effect is a phenomenon that occurs in many complex systems, such as weather. Predictions can be subject to significant uncertainty due to small errors in the initial state.

- **Non-Stationarity**

As uncertainty changes in dynamic systems, adaptive assimilation techniques are necessary.

5.1.7. Real-time Requirements

In some data assimilation applications, such as weather forecasting and disaster management, real-time or near-real-time processing is required. There is a possibility that deep learning models may introduce latency that prevents them from meeting these requirements.

- **High Computational Demands**

Real-time requirements for complex tasks, such as real-time image processing, video analysis, or simulations, often require significant computational resources, making them difficult to achieve on standard hardware.

- **Data Throughput**

In order to maintain real-time processing, real-time systems that deal with high volumes of data, such as sensor networks or streaming data analytics, must manage data efficiently.

- **System Variability**

There is a significant amount of variability in the execution times of real-time systems as a result of factors such as re-source contention, varying workloads, or hardware failure.

- **Predictability Challenge**

The challenge of ensuring consistent, predictable performance in the face of such variability is significant.

- **Concurrency and Synchronization**

Managing concurrency: To avoid conflicts and meet deadlines, real-time systems must efficiently handle synchronization and resource allocation.

- **Fault Tolerance**

It may be necessary to include redundant components in real-time systems to en-sure that they remain functional in the event of hardware or software failures.

- **Safety-Critical Concerns**

Real-time systems must meet rigorous certification standards in safety-critical domains like aviation and healthcare, adding complexity and cost.

5.2. Suggestions and prospects

The research above indicates that although deep learning approaches have great potential for assimilation of data, there are a number of limitations and challenges that are hindering their development. There is an urgent need to overcome these limitations and challenges at the present time. In this section, some suggestions are provided.

5.2.1. Suggestions for data gaps

Data deficiencies in the context of deep learning-based data assimilation are an active area of research.

- **Imputation Techniques**

Researchers can focus on developing advanced imputation techniques, including deep learning-based methods, to fill gaps in observational data. As a result, missing values must be predicted while taking into account spatial and temporal dependencies.

- **Synthetic Data Generation**

It is possible for researchers to use generative models, such as GANs and VAEs, to create synthetic data that complements observational data. It is possible to mitigate deficiencies in training data by using synthetic data.

- **Deep Learning for Quality Control**

Study the use of deep learning algorithms to automate quality control procedures for observational data. Data errors, outliers, and biases can be identified and corrected using these algorithms.

- **Bias Correction Models**

Develop deep learning models that can effectively correct biases in observational data, especially when dealing with historical data or data from different sources.

- **Multimodal Models**

Develop deep learning models that can effectively fuse and integrate data from multiple sources and modalities. There are several techniques for combining data with

different resolutions, scales, and types (e.g., satellite imagery, ground-based measurements, remote sensing).

- **Transfer Learning**

Research transfer learning techniques that can be used to adapt models that have been trained on one dataset or modality to new datasets or domains that have a limited amount of labeled data.

- **Uncertainty-Aware Models**

Developing deep learning models that explicitly quantify and propagate uncertainty from observational data to model predictions. In this context, Bayesian deep learning and probabilistic modeling can be very useful.

- **Data-Driven Uncertainty Estimation**

Explore the possibility of using deep generative models to estimate the uncertainty in observational data based on historical records and sensor characteristics.

- **Sparse Data Models**

Develop deep learning architectures that are specifically designed to handle sparse data, whether due to limitations in data collection or inequalities in spatial and temporal coverage.

- **Active Learning**

Examine active learning techniques that reduce data sparsity and improve model performance by strategically selecting observations.

- **Optimal Data Collection**

Optimize data collection strategies to maximize the informativeness of observational data while minimizing costs.

- **Edge Computing**

Investigate the feasibility of deploying lightweight deep learning models on edge devices for the assimilation of real-time data in resource-constrained environments.

- **Interpretability Methods**

Improve the interpretability of deep learning models in the context of data assimilation by developing techniques and tools. Understanding how models make decisions and ensuring physical consistency are crucial.

- **Hybrid Models**

Analyze hybrid models that combine deep learning components with traditional methods of data assimilation that can be interpreted.

5.2.2. Suggestions for *overfitting* and generalization

To advance these research directions and address the challenges of overfitting and generalization, researchers, domain experts, and data assimilation practitioners must collaborate. It is imperative that the reliability and robustness of deep learning-based approaches are improved to achieve accurate predictions and simulations in complex, dynamic systems, since data assimilation continues to play a critical role in a variety of scientific and environmental applications.

- **Data Augmentation Techniques**

Develop techniques for enhancing observational data with realistic variations, such as perturbing measurements or simulating missing data. By doing so, it is possible to create a more diverse training dataset, which can reduce the likelihood of overfitting.

- **Regularization Strategies**

Investigate novel regularization techniques that are tailored to data assimilation tasks. There may be adaptive regularization schemes that automatically adjust regularization strength according to the characteristics of the data.

- **Ensemble Learning**

Extend ensemble learning approaches by incorporating diversity-enhancing techniques, such as bootstrapping and different architectures, to further reduce overfitting and improve model robustness.

- **Bayesian Deep Learning**

Investigate Bayesian deep learning methods that provide principled approaches for quantifying and managing model uncertainty. To account for epistemic uncertainty associated with limited data, Bayesian neural networks can be applied to data assimilation tasks.

- **Domain Adaptation**

Design domain adaptation techniques that explicitly address the challenges involved in adapting deep learning models to new regions, environmental conditions, or time periods. In the absence of training data, domain adaptation may help models generalize more effectively.

- **Covariate Shift Detection**

Methods for detecting and correcting covariate shifts in data assimilation should be investigated. In real-time, the detection of shifts in the distribution of data can assist models in adapting and maintaining their generalization abilities.

- **Multi-Source Data Assimilation**

Study multi-source data assimilation approaches that leverage data from multiple sources, including different regions and time periods. Using diverse data sources can improve the model's ability to generalize under a variety of conditions.

- **Spatiotemporal Consistency**

Develop techniques for ensuring spatiotemporal consistency in data assimilation models. In order to improve generalization to new conditions, model predictions should be consistent with the physical laws governing the system.

- **Interpolation and Extrapolation**

Analyze methods for improving the interpolation and extrapolation capabilities of the model. Robust generalization requires models that can accurately predict values between and beyond observation points.

- **Model Explainability and Uncertainty**

Improve model interpretability and uncertainty representation in deep learningbased data assimilation. Assimilation results can be trusted if transparent models and well-calibrated uncertainty estimates are used.

- **Incremental Learning**

Explore incremental learning techniques that enable models to continuously adapt and update their knowledge as

new data becomes available. Assimilation scenarios with evolving conditions are particularly relevant to this.

5.2.3. Suggestions *for* interpretability

Collaboration between researchers, practitioners, and policymakers will be required to address these challenges and advance the field of deep learning interpretability. The adoption of deep learning models across industries continues to grow, making mutually explainable AI systems increasingly important.

- **Interpretation Metrics**

Identify metrics that can be used to evaluate the interpretability of deep learning models for specific tasks and domains. Metrics such as fidelity, consistency, and user satisfaction with explanations may be considered.

- **Task-Oriented Interpretations**

Techniques for generating task-specific interpretations tailored to the needs of end users. It may be necessary to provide different levels of detail and context in explanations for different tasks.

- **Model-Agnostic Interpretability**

Develop model-agnostic interpretability techniques that can be applied to various deep learning architectures. Methods that are model-agnostic, such as LIME and SHAP, aim to provide insights into black-box models.

- **Uncertainty Estimation**

By extending model-agnostic techniques, users will be able to receive uncertainty estimates along with interpretations, providing them with insight into the reliability of the model.

- **Interpretable Model Architectures**

Develop inherently interpretable model architectures. Neural network designs that maintain transparency and reveal decision-making processes are included in this category.

- **Interpretable Pretraining**

Investigate methods for pretraining deep models with interpretable objectives or representations. As a result, models can be fine-tuned for specific purposes with greater interpretability.

5.2.4. *Suggestions* for *model* complexity

In the field of deep learning and deep learning, balancing the complexity of models with their limitations is a continuing challenge. Here are some suggestions for addressing these limitations and advancing research in the field:

- **Advanced Regularization Techniques**

Develop advanced regularization techniques for controlling overfitting in complex models. Regularization strength can be adaptively adjusted during training based on the performance of the model.

- **Structured Pruning**

Investigate structured pruning methods that reduce computational demands and preserve the overall structure of complex models when removing entire neurons or subnetworks.

- **Dynamic Ensembling**

Optimize complexity and accuracy trade-offs by dynamically adapting ensemble members during inference.

- **Sparse and Efficient Architectures**

Develop sparse and efficient model architectures that maintain high performance while reducing computational and memory demands.

5.2.5. *Suggestions for domain specificity*

It is essential to explore more detailed solutions and research directions in order to overcome the limitations of domain-specific models and harness their full potential:

- **Domain Adaptation Techniques**

Develop advanced domain adaptation techniques that allow domain-specific models to transfer knowledge and adapt more effectively to related domains.

- **Meta-Learning**

Analyze meta-learning approaches that enable domain-specific models to learn how to adapt quickly to new domains, even with limited data.

- **Data Augmentation and Synthesis**

Investigate the use of generative models, such as GANs (Generative Adversarial Networks), to create synthetic data that closely resembles the domain of interest, thus addressing the issue of data scarcity.

5.2.6. *Suggestions* for assimilation of uncertainty

We propose the following points in order to address these limitations and advance the uptake of uncertainty:

- **Advanced Uncertainty Models**

In order to better capture the characteristics of real-world uncertainty, develop and implement advanced uncertainty models, such as non-Gaussian distributions and heavy-tailed distributions.

- **Data-Driven Techniques**

Using machine learning and data-driven approaches to assimilate uncertainty can assist in handling complex, high-dimensional uncertainty sources.

- **Robust Parameter Estimation**

Enhance techniques for estimating uncertain model parameters, taking uncertainty into account.

- **Error Propagation Studies**

Investigate how uncertainties in various components of the assimilation process affect the final predictions through comprehensive error propagation studies.

- **Robust Assimilation**

Develop techniques for robust assimilation that can accommodate multiple sources of uncertainty and model errors.

5.2.7. Suggestions *for* real-time requirements

To address these limitations, this paper makes the following recommendations:

- **Real-time Operating Systems (RTOS)**

Develop real-time operating systems and middleware that enable predictable and efficient task scheduling.

- **Hardware Acceleration**

To meet computational demands, investigate the use of field-programmable gate arrays (FPGAs) and graphics processing units (GPUs).

- **Real-time Scheduling Algorithms**

Implement advanced real-time scheduling algorithms that can handle complex task dependencies and system dynamics.

- **Energy-Efficient Computing**

Research energy efficient scheduling algorithms that balance real-time requirements with power-saving measures.

- **Real-time Debugging and Monitoring**

Develop debugging tools specifically designed for real-time systems that do not affect their timing constraints.

- **Edge Computing and Edge AI**

Investigate edge computing and edge AI approaches to offload processing from centralized systems and to meet real-time requirements in distributed environments.

Author Contributions

Conceptualization, Y.Z. and D.Z; methodology,Y.Z. and D.Z.; software, D.Z.; vali-dation, D.Z. and Y.Z.; formal analysis, Y.Z. and D.Z; investigation, H.J. and D.Z; resources, D.Z.; data curation, Y.M.; writing-original draft preparation, Y.Z.; writing-review and editing, D.Z.; visualization, Y.M.; supervision, D.Z.; funding acquisition, D.Z. and J.H. All authors have read and agreed to the published version of the manuscript.

Acknowledgements

References

[1] Bi, Kaifeng, et al, Accurate medium-range global weather forecasting with 3D neural networks, Nature (2023) 1–6.

[2] Ling, Fenghua et al., Multi-task machine learning improves multi-seasonal prediction of the Indian Ocean Dipole." Nature Communications 13 (2022) 7681.

[3] Tiggeloven, Timothy, et al., Exploring deep learning capabilities for surge predictions in coastal areas, Scientific reports 11 (2021) 17224.

[4] Zhang, Yi, Dapeng Zhang, and Haoyu Jiang, Review of Challenges and Opportunities in Turbulence Modeling: A Comparative Analysis of Data-Driven Machine Learning Approaches, Journal of Marine Science and Engineering 11 (2023) 1440.

[5] Duraisamy, Karthik, Gianluca Iaccarino, and Heng Xiao, Turbulence modeling in the age of data, Annual review of fluid mechanics 51 (2019) 357–377.

[6] Zhang, Yi, Dapeng Zhang, and Haoyu Jiang, A Review of Offshore Wind and Wave Installations in Some Areas with an Eye towards Generating Economic Benefits and Offering Commercial Inspiration, Sustainability 15 (2023) 8429.

[7] Robertson, Bryson, Jessica Bekker, and Bradley Buckham. "Renewable integration for remote communities: Comparative allowable cost analyses for hydro, solar and wave energy, Applied Energy 264 (2020) 114677.

[8] Senchenko, Victor, et al., Technical automation tools for high-precision navigating of sea and river ships." International Scientific Conference on Architecture and Construction. Singapore: Springer Nature Singapore, 2020.

[9] Minami, Makiko, et al., Development of the Comprehensive Simulation System for Autonomous Ships, Journal of Physics: Conference Series, 2311 (2022).

[10] Liu, Chenguang, et al., Human–machine cooperation research for navigation of maritime autonomous surface ships: A review and consideration, Ocean Engineering 246 (2022) 110555.

[11] Evensen, Geir, Femke C. Vossepoel, and Peter Jan van Leeuwen. Data assimilation fundamentals: A unified formulation of the state and parameter estimation problem. Springer Nature, 2022.

[12] Zhang, Shaoqing, et al., Coupled data assimilation and parameter estimation in coupled ocean–atmosphere models: a review." Climate Dynamics 54 (2020) 5127–5144.

[13] Rogers, Cassandra, and Chris Tingwell. Forecast sensitivity to the assimilation of observational data-two case studies for Australia. No. EGU23-14259. Copernicus Meetings, 2023.

[14] Chattopadhyay, Ashesh, et al., Deep learning-enhanced ensemble-based data assimilation for high-dimensional nonlinear dynamical systems, Journal of Computational Physics 477 (2023) 111918.

[15] Baracchini, Theo, et al., Data assimilation of in situ and satellite remote sensing data to 3D hydrodynamic lake models: a case study using Delft3D-FLOW v4. 03 and OpenDA v2. 4. Geoscientific Model Development 13 (2020) 1267–1284.

[16] Quetin, Gregory R., et al., Carbon flux variability from a relatively simple ecosystem model with assimilated data is consistent with terrestrial biosphere model estimates, Journal of Advances in Modeling Earth Systems 12 (2020) e2019MS001889.

[17] Dong, Shi, Ping Wang, and Khushnood Abbas, A survey on deep learning and its applications, Computer Science Review 40 (2021) 100379.

[18] Niu, Zhaoyang, Guoqiang Zhong, and Hui Yu, A review on the attention mechanism of deep learning, Neurocomputing 452 (2021) 48–62.

[19] Janiesch, Christian, Patrick Zschech, and Kai Heinrich. ,Machine learning and deep learning, Electronic Markets 31.3 (2021) 685–695.

[20] Gupta, B., Rahmi, S. and Yang, Y., 2007. A novel roll-back mechanism for performance enhancement of asynchronous checkpointing and recovery. Informatica, 31(1).

[21] Jahanbakht, Mohammad, et al. "Internet of underwater things and big marine data analytics—a comprehensive survey." IEEE Communications Surveys & Tutorials 23.2 (2021): 904-956.

[22] N. Marhamati et al., Integration of Z-numbers and Bayesian decision theory: A hybrid approach to decision making under uncertainty and imprecision, Applied Soft Computing, 72 (2018) 273–290, https://doi.org/10.1016/j.asoc.2018.07.053.

[23] Manasrah, Riyad, et al., Physical and chemical properties of seawater during 2013–2015 in the 400 m water column in the

northern Gulf of Aqaba, Red Sea, Environmental monitoring and assessment 192 (2020): 1-16.

[24] Khorasani, E.S., Patel, P., Rahimi, S. et al. An inference engine toolkit for computing with words. J Ambient Intell Human Comput **4**, 451–470 (2013). https://doi.org/10.1007/s12652-012-0137-8

[25] S. Neupane et al., Explainable Intrusion Detection Systems (X-IDS): A Survey of Current Methods, Challenges, and Opportunities, IEEE Access, 10 (2022) 112392–112415, doi: 10.1109/ACCESS.2022.3216617.

[26] Wu, Peng, et al., Inversion of deep-water velocity using the Munk formula and the seabed reflection traveltime: An inversion scheme that takes the complex seabed topography into account, IEEE Transactions on Geoscience and Remote Sensing (2023).

[27] Wang, Bing, et al., Seabed features associated with cold seep activity at the Formosa Ridge, South China Sea: Integrated application of high-resolution acoustic data and photomosaic images, Deep Sea Research Part I: Oceanographic Research Papers 177 (2021) 103622.

[28] Tang, Qiuhua, et al., Deep-sea seabed sediment classification using finely processed multibeam backscatter intensity data in the southwest Indian ridge. Remote Sensing 14 (2022) 2675.

[29] Adland, Roar, et al., The value of meteorological data in marine risk assessment, Reliability Engineering & System Safety 209 (2021) 107480.

[30] Freeman, Eric, et al., The international comprehensive ocean-atmosphere data set–meeting users needs and future priorities, Frontiers in Marine Science 6 (2019) 435.

[31] Canonico, Gabrielle, et al., Global observational needs and resources for marine biodiversity, Frontiers in Marine Science 6 (2019) 367.

[32] Zhao, Qianshuo, et al. "Where marine protected areas would best represent 30% of ocean biodiversity." Biological Conservation 244 (2020) 108536.

[33] Sunagawa, Shinichi, et al. "Tara Oceans: towards global ocean ecosystems biology." Nature Reviews Microbiology 18.8 (2020) 428-445.

[34] Brito-Morales, Isaac, et al., Climate velocity reveals increasing exposure of deep-ocean biodiversity to future warming, Nature Climate Change 10.6 (2020) 576-581.

[35] Groom, Steve, et al., Satellite ocean colour: current status and future perspective, Frontiers in Marine Science 6 (2019) 485.

[36] Watanabe, Jun-Ichiro, Yang Shao, and Naoto Miura, Underwater and airborne monitoring of marine ecosystems and debris, Journal of Applied Remote Sensing 13 (2019) 044509.

[37] Loeb, Norman G., et al., Satellite and ocean data reveal marked increase in Earth's heating rate, Geophysical Research Letters 48 (2021) e2021GL093047.

[38] Wang, Hengyu, et al., A New Wave Energy Converter for Marine Data Buoy, IEEE Transactions on Industrial Electronics 70.2 (2022) 2076-2084.

[39] Perez, Renellys C., et al., Oceanographic buoys: Providing ocean data to assess the accuracy of variables derived from satellite measurements, Field Measurements for Passive Environmental Remote Sensing. Elsevier, (2023) 79–100.

[40] Xu, Ruijiang, et al., Recent progress on wave energy marine buoys, Journal of Marine Science and Engineering 10.5 (2022): 566.

[41] Knight, Philip J., et al., A low-cost GNSS buoy platform for measuring coastal sea levels, Ocean Engineering 203 (2020) 107198.

[42] Kremser, Stefanie, et al.,Southern Ocean cloud and aerosol data: a compilation of measurements from the 2018 Southern Ocean Ross Sea Marine Ecosystems and Environment voyage, Earth System Science Data 13 (2021) 3115–3153.

[43] Di Luccio, Diana, et al., Coastal marine data crowdsourcing using the Internet of Floating Things: Improving the results of a water quality model, IEEE Access 8 (2020) 101209–101223.

[44] Neira, Javier, et al., Review on unmanned underwater robotics, structure designs, materials, sensors, actuators, and navigation control, Journal of Robotics 2021 (2021) 1–26.

[45] Sánchez, Pedro José Bernalte, Mayorkinos Papaelias, and Fausto Pedro García Márquez, Autonomous underwater vehicles: Instrumentation and measurements, IEEE Instrumentation & Measurement Magazine 23.2 (2020) 105-114.

[46] Angryk, R., Kołodziej, K., Fiedorowicz, I., Paprzycki, M., Cobb, M., Ali, D. and Rahimi, S., 2001. Development of a travel support system based on intelligent agent technology. In Proceedings of the PIONIER 2001 Conference (S. Niwiński ed.) Technical University of Poznań Press, Poznań, Poland (pp. 243-255).

[47] Keisling, Clarissa, et al., Low concentrations and low spatial variability of marine microplastics in oysters (Crassostrea virginica) in a rural Georgia estuary, Marine pollution bulletin 150 (2020) 110672.

[48] Jiang, Li-Qing, et al., Surface ocean pH and buffer capacity: past, present and future, Scientific reports 9.1 (2019) 18624.

[49] Mansui, J., et al., Predicting marine litter accumulation patterns in the Mediterranean basin: Spatio-temporal variability and comparison with empirical data." Progress in Oceanography 182 (2020) 102268.

[50] Skalska, Karolina, et al., Riverine microplastics: Behaviour, spatio-temporal variability, and recommendations for standardised sampling and monitoring, Journal of Water Process Engineering 38 (2020): 101600.

[51] Mirimin, Luca, et al., Don't catch me if you can–Using cabled observatories as multidisciplinary platforms for marine fish community monitoring: an in situ case study combining Underwater Video and environmental DNA data, Science of the Total Environment 773 (2021) 145351.

[52] Huang, Mingfeng, et al., An AUV-assisted data gathering scheme based on clustering and matrix completion for smart ocean, IEEE Internet of Things Journal 7 (2020) 9904–9918.

[53] Kent, Elizabeth C., et al., Observing requirements for long-term climate records at the ocean surface, Frontiers in Marine Science 6 (2019) 441.

[54] Tang, Meng, Yimin Liu, and Louis J. Durlofsky, A deep-learning-based surrogate model for data assimilation in dynamic subsurface flow problems, Journal of Computational Physics 413 (2020) 109456.

[55] Arcucci, Rossella, et al., Deep data assimilation: integrating deep learning with data assimilation, Applied Sciences 11 (2021) 1114.

[56] Brajard, Julien, et al. "Combining data assimilation and machine learning to infer unresolved scale parametrization." Philosophical Transactions of the Royal Society A 379.2194 (2021): 20200086.

[57] Kingma, Diederik P., and Max Welling, An introduction to variational autoencoders, Foundations and Trends® in Machine Learning 12.4 (2019): 307-392.

[58] Grooms, Ian, Analog ensemble data assimilation and a method for constructing analogs with variational autoencoders, Quarterly Journal of the Royal Meteorological Society 147 (2021) 139–149.

[59] Canchumuni, Smith WA, Alexandre A. Emerick, and Marco Aurélio C. Pacheco, Towards a robust parameterization for conditioning facies models using deep variational autoencoders and ensemble smoother, Computers & Geosciences 128 (2019) 87–102.

[60] Gupta, B., Rahimi, S. and Liu, Z., 2006. A new high performance checkpointing approach for mobile computing

systems. IJCSNS International Journal of Computer Science and Network Security, 6 (2006)95–104.

[61] Cheng, Sibo, et al, Machine learning with data assimilation and uncertainty quantification for dynamical systems: a review, IEEE/CAA Journal of Automatica Sinica 10 (2023) 1361–1387.

[62] Zhang, Kai, et al, Multi-source information fused generative adversarial network model and data assimilation based history matching for reservoir with complex geologies, Petroleum Science 19 (2022) 707–719.

[63] Yu, Yong, et al., A review of recurrent neural networks: LSTM cells and network architectures., Neural computation 31.7 (2019) 1235–1270.

[64] Weerakody, Philip B., et al., A review of irregular time series data handling with gated recurrent neural networks, Neurocomputing 441 (2021) 161–178.

[65] Wang, Yunbo, et al., Predrnn: A recurrent neural network for spatiotemporal predictive learning, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2022) 2208–2225.

[66] Penny, Stephen G., et al., Integrating recurrent neural networks with data assimilation for scalable data-driven state estimation, Journal of Advances in Modeling Earth Systems 14 (2022): e2021MS002843.

[67] Cheng, Sibo, and Mingming Qiu, Observation error covariance specification in dynamical systems for data assimilation using recurrent neural networks, Neural Computing and Applications 34 (2022) 13149–13167.

[68] Casas, César Quilodrán, et al. "A reduced order deep data assimilation model." Physica D: Nonlinear Phenomena 412 (2020): 132615.

[69] Tsuyuki, Tadashi, and Ryosuke Tamura. "Nonlinear data assimilation by deep learning embedded in an ensemble Kalman filter." Journal of the Meteorological Society of Japan. Ser. II 100.3 (2022): 533-553.

[70] Arcucci, Rossella, et al. "Deep data assimilation: integrating deep learning with data assimilation." Applied Sciences 11.3 (2021): 1114.

[71] Zhang, Yonggang, et al. "The adoption of deep neural network (DNN) to the prediction of soil liquefaction based on shear wave velocity." Bulletin of Engineering Geology and the Environment 80 (2021): 5053-5060.

[72] Wan, Zhi, et al., Optimization of vascular structure of self-healing concrete using deep neural network (DNN), Construction and Building Materials 364 (2023) 129955.

[73] Cuong-Le, Thanh, et al., A novel version of grey wolf optimizer based on a balance function and its application for hyperparameters optimization in deep neural network (DNN) for structural damage identification, Engineering Failure Analysis 142 (2022) 106829.

[74] Deng, Zhiwen, Chuangxin He, and Yingzheng Liu, Deep neural network-based strategy for optimal sensor placement in data assimilation of turbulent flow, Physics of Fluids 33.2 (2021).

[75] Misaka, Takashi, Image-based fluid data assimilation with deep neural network, Structural and Multidisciplinary Optimization 62 (2020): 805–814.

[76] Y. Zhang, D. Zhang, Y. Zhang, Y. Xie, B. xie, H. Jiang, A Comprehensive Review of Simulation Software and Experiemntal Modeling on Exploring Marine Collision Analysis, ENG Transactions 4 (2023) 1–7, Article ID: 2869.

[77] Pawar, Suraj, and Omer San, Data assimilation empowered neural network parametrizations for subgrid processes in geophysical flows, Physical Review Fluids 6 (2021) 050501.

[78] Khorasani, E.S., Rahimi, S., Patel, P. and Houle, D., 2011, September. Cwjess: Implementation of an expert system shell for computing with words. In 2011 Federated Conference on Computer Science and Information Systems (FedCSIS) (pp. 33–39). IEEE.

[79] Li, Zewen, et al., A survey of convolutional neural networks: analysis, applications, and prospects, IEEE transactions on neural networks and learning systems (2021).

[80] Zhou, Ding-Xuan. Theory of deep convolutional neural networks: Downsampling." Neural Networks 124 (2020) 319–327.

[81] Rahimi, S., Carver, N., Petry, F. (2005). A Multi-Agent Architecture for Distributed Domain-Specific Information Integration. In: Ladner, R., Petry, F.E. (eds) Net-Centric Approaches to Intelligence and National Security. Springer, Boston, MA. https://doi.org/10.1007/0-387-26176-1_7

[82] Lei, T., Luo, C., Ball, J.E. and Rahimi, S. A graph-based ant-like approach to optimal path planning. In 2020 IEEE congress on evolutionary computation (CEC) IEEE (2020) 1–6.

[83] Scott, K. Andrea, Linlin Xu, and Homa Kheyrollah Pour, Retrieval of ice/water observations from synthetic aperture radar imagery for use in lake ice data assimilation." Journal of Great Lakes Research 46 (2020) 1521–1532.

[84] Han, Jingye, et al., Rice yield estimation using a CNN-based image-driven data assimilation framework, Field Crops Research 288 (2022) 108693.

[85] Gao, Han, Luning Sun, and Jian-Xun Wang, Super-resolution and denoising of fluid flow using physics-informed convolutional neural networks without high-resolution labels. Physics of Fluids 33 (2021).

[86] Gui, Jie, et al., A review on generative adversarial networks: Algorithms, theory, and applications, IEEE transactions on knowledge and data engineering 35.4 (2021) 3313-3332.

[87] Jabbar, Abdul, Xi Li, and Bourahla Omar, A survey on generative adversarial networks: Variants, applications, and training. ACM Computing Surveys (CSUR) 54 (2021) 1–49.

[88] Saxena, Divya, and Jiannong Cao, Generative adversarial networks (GANs) challenges, solutions, and future directions, ACM Computing Surveys (CSUR) 54.3 (2021) 1–42.

[89] Bao, Jichao, Liangping Li, and Fleford Redoloza, Coupling ensemble smoother and deep learning with generative adversarial networks to deal with non-Gaussianity in flow and transport data assimilation, Journal of Hydrology 590 (2020) 125443.

[90] Zhang, Kai, et al., Multi-source information fused generative adversarial network model and data assimilation based history matching for reservoir with complex geologies, Petroleum Science 19 (2022) 707–719.

[91] Silva, Vinicius LS, et al., Data assimilation predictive GAN (DA-PredGAN) applied to a spatio-temporal compartmental model in epidemiology, Journal of Scientific Computing 94.1 (2023) 25.

[92] Silva, Vinicius LS, et al., Data assimilation predictive GAN (DA-PredGAN) applied to a spatio-temporal compartmental model in epidemiology, Journal of Scientific Computing 94 (2023) 25.

[93] Niu, Zhaoyang, Guoqiang Zhong, and Hui Yu, A review on the attention mechanism of deep learning, Neurocomputing 452 (2021) 48–62.

[94] Khan, Salman, et al., Transformers in vision: A survey, ACM computing surveys (CSUR) 54 (2022) 1–41.

[95] Gao, Chuan, Lu Zhou, and Rong-Hua Zhang, A Transformer-Based Deep Learning Model for Successful Predictions of the 2021 Second-Year La Niña Condition, Geophysical Research Letters 50.12 (2023): e2023GL104034.

[96] Li, Klin, and Yutong Lu, A Transformer-Based Framework for Parameter Learning of a Land Surface Hydrological Process Model, Remote Sensing 15 (2023) 3536.

[97] Jun, Jimin, and Hong Kook Kim, Informer-Based Temperature Prediction Using Observed and Numerical Weather Prediction Data, Sensors 23.16 (2023) 7047.

[98] Rahimi, S., Gandy, L. and Mogharreban, N., A web-based high-performance multicriteria decision support system for medical diagnosis, International Journal of Intelligent Systems 22 (2007) 1083–1099.

[99] Mohammed, Ammar, and Rania Kora, A comprehensive review on ensemble deep learning: Opportunities and challenges, Journal of King Saud University-Computer and Information Sciences (2023).

[100] Gupta, B., Rahimi, S., Rias, R.A. and Bangalore, G.,. A low-overhead non-block checkpointing algorithm for mobile computing environment. In Advances in Grid and Pervasive Computing: First International Conference, GPC 2006, Taichung, Taiwan, May 3-5, 2006. Springer Berlin Heidelberg.Proceedings 1 (2006) 597–608.

[101] Yang, L. Minah, and Ian Grooms. Machine learning techniques to construct patched analog ensembles for data assimilation. Journal of Computational Physics 443 (2021) 110532.

[102] Laleh, A. and Shahram, R., December. Analyzing Facebook activities for personality recognition. In 2017 16th IEEE international conference on machine learning and applications (ICMLA) IEEE (2017) 960–964.

[103] Chattopadhyay, Ashesh, et al, Deep learning-enhanced ensemble-based data assimilation for high-dimensional nonlinear dynamical systems, Journal of Computational Physics 477 (2023) 111918.

[104] Z. Khashroum, H. Rahimighazvini, M. Bahrami, Applications of Machine Learning in Power Electronics: A Specialization on Convolutional Neural Networks, ENG Transactions 4 (2023) 1–5, Article ID: 2866.

[105] S. B. Ramezani et al., A novel compartmental model to capture the nonlinear trend of COVID-19, Computers in Biology and Medicine, 134 (2021) 104421, https://doi.org/10.1016/j.compbiomed.2021.104421.

[106] T. S. Tabrizi et al., Towards a patient satisfaction based hospital recommendation system, 2016 International Joint Conference on Neural Networks (IJCNN), Vancouver, BC, Canada, (2016) 131–138, doi: 10.1109/IJCNN.2016.7727190.

[107] Y. Luo, J. Yan, D. Zhang, Stochastic Response Analysis of Two Vibration Systems with Impact Interactions, ENG Transactions 4 (2023) 1–8, Article ID: 2868.