IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 14, No. 2, April 2025, pp. 867~883
ISSN: 2252-8938, DOI: 10.11591/ijai.v14.i2.pp867-883

Bridging biosciences and deep learning for revolutionary discoveries: a comprehensive review

Usman Tariq[1], Irfan Ahmed[2], Muhammad Attique Khan[3], Ali Kashif Bashir[4]
[1]Department of Management Information System, College of Business Administration, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia
[2]College of Engineering, Virginia Commonwealth University, Richmond, United States of America
[3]Department of Computer Science, HITEC University, Taxila, Pakistan
[4]School of Computing, Mathematics and Digital Technology, Manchester Metropolitan University, Manchester, United Kingdom

Article Info

Article history:

Received Mar 24, 2024
Revised Nov 18, 2024
Accepted Nov 24, 2024

Keywords:

Artificial intelligence
Deep learning
Neural networks
Transformed analysis
Variant calling

ABSTRACT

Deep learning (DL), a pivotal artificial intelligence (AI) innovation, has dramatically transformed biosciences, aligning with the surge in complex data volumes to foster notable progress across disciplines such as genomics, genetics, and drug discovery. DL's precision and efficiency outmatch conventional methods, propelling advancements in biomedical imaging and disease marker identification. Despite its success, DL's integration into broader bioscience areas encounters hurdles including data scarcity, interpretability challenges, computational demands, and the necessity for ethical and regulatory considerations. Overcoming these obstacles is vital for DL to achieve its transformative potential fully. This review explores into DL's expanding role in biosciences, critically examining areas ripe for DL application and highlighting underexplored opportunities. It provides an insightful analysis of the algorithms that form the backbone of DL in biosciences, offering a thorough understanding of their capabilities. Ultimately, this paper aims to equip biotechnologists and researchers with the knowledge to leverage DL effectively, thereby enhancing the analysis of complex bioscience data and contributing to the field's future advancements.

Corresponding Author:

Usman Tariq
Department of Management Information System, College of Business Administration
Prince Sattam Bin Abdulaziz University
Al-Kharj 16278, Saudi Arabia
Email: u.tariq@psau.edu.sa

1. INTRODUCTION

Deep learning (DL) advent has markedly reshaped the field of computational sciences, distinguishing itself from conventional machine learning with its adept handling of complex data structures. This transformation is attributed to a blend of algorithmic enhancements and a surge in computational capacity, notably in graphics processing unit (GPUs) [1], coupled with the availability of large datasets. Such developments have spurred advancements in areas like natural language processing and computer vision. Concurrently, the foray of DL into biosciences corresponds with the influx of diverse, large-scale biological data, driven by novel technologies including next-generation sequencing. Renowned for its proficiency in feature extraction and pattern recognition within massive datasets, DL is pivotal in sectors such as genomics and disease analysis, signifying a shift to-wards a more data-focused research paradigm. This integration heralds new pathways in personalized medicine and introduces issues like data diversity, infrastructural needs, and ethical dilemmas.

Understanding DL's application in biosciences entails familiarizing oneself with its basic concepts and frameworks, along with their pros and cons. Neural networks, e.g. convolutional neural networks (CNNs) [2] and recurrent neural networks (RNNs) [3], are fundamental to DL, each being specialized for certain data types and applications. CNNs are crucial for examining spatial data, such as medical imagery, whereas RNNs are significant in sequencing genetic data. These networks excel in identifying complex patterns through their layered and nonlinear processing abilities. DL shines in its ability to manage extensive, unstructured datasets, automate feature extraction, and continuously improve with additional data. Nonetheless, DL models can be opaque, particularly in clinical settings, and rely heavily on high-quality data and substantial computational power. Recognizing these aspects is essential for the practical application of DL in biosciences [4], [5].

Our comprehensive research reveals that the biosciences sector is amid a data revolution, brimming with challenges and opportunities. Swift technological progress, especially in sequencing and imaging, has precipitated a dramatic rise in data volume. This surge offers immense potential for discovery but also poses significant hurdles in data handling and analysis. In this context, DL stands out as a potent solution for analysing complex, voluminous datasets to reveal concealed insights. DL is not just managing this data deluge but is also converting it into a source of scientific innovation and breakthroughs.

Likewise, DL is making significant strides in biosciences, impacting areas from genomics to disease diagnostics. It is instrumental in refining the accuracy of gene function identification in genomics. Similarly, DL is vital for elucidating protein structures and functions, thus unravelling complex biological processes. Its growing role in drug discovery and development is apparent, hastening the identification of new medical treatments. It also enhances the precision of biomedical imaging and plays a key role in biomarker discovery, leading to earlier and more accurate disease detection. The extensive range of these applications underscores DL's potential to revolutionize bioscience research and healthcare.

This review thoroughly investigates the key areas where DL finds application within biosciences, taking a structured approach to analyze various research studies and emerging trends. It aims to provide a detailed understanding of how DL is applied across different domains, such as genomics and drug discovery, while carefully examining the significance of each application and the methods utilized. The expected outcome is a deep understanding of how DL influences biosciences, underscoring its role in advancing research and practical use, and laying the groundwork for future progress in the field.

This review paper is organized as follows: section 2 dives into genomics and genetics, presenting the latest advancements and challenges in this area. Section 3 explores the transformative role of these intelligent deep-learning algorithms in drug discovery and development, highlighting key methodologies and breakthroughs. Section 4 addresses the complexities and potential of patient stratification, showcasing its impact on personalized medicine. Section 5 offers an in-depth analysis of various models, detailing their functionality, advantages, disadvantages, and diverse applications. Section 6 shifts focus to the synergy of quantum computing and intelligent algorithms, elaborating on artificial intelligence (AI) roles in pandemic response, agricultural studies, neuroinformatics, and ecosystem modelling, emphasizing quantum computing's enhancement of analytical capabilities. Section 7 critically evaluates ethical and regulatory considerations, including privacy, bias, and compliance issues. Section 8 anticipates prospective trends, forecasting future directions and innovations. Finally, section 9 summarizes key insights and findings, offering reflections on the path ahead in the domain. Each section aims to provide a comprehensive understanding of the technical intricacies and broader implications of these technologies in advancing research and applications.

2. DEEP LEARNING IN GENOMICS AND GENETICS

DL is transforming genomics and genetics, making complex tasks like genome assembly and annotation more efficient. Utilizing algorithms such as DeepVariant and genome analysis toolkit (GATK) [6], DL enhances the accuracy in assembling and annotating genomic sequences. It also improves variant calling and genotyping, essential for identifying dis-ease-linked genetic variations, by distinguishing true variants from sequencing errors.

In gene expression analysis, DL tools like DeepGene [7] and Seq2Gene [8] are instrumental in interpreting gene expression patterns, enhancing our understanding of cellular processes. DL also aids in predicting regulatory elements, which is crucial for studying gene regulation and complex diseases like cancer. Besides, genome-wide association studies (GWAS) [9] have become more potent with DL, enabling the analysis of large datasets to uncover new genetic-disease associations. In personalized medicine and precision genomics, DL is key, with tools such as DeepGenome and PrecisionMed analysing patient genetic data for tailored treatments.

3. DEEP LEARNING IN DRUG DISCOVERY AND DEVELOPMENT

Our exclusive research investigation showed that DL plays a vital role in drug discovery and development [10]. For instance, i) in virtual screening and lead identification, DL models (e.g. DL target interaction based on CNN, deep learning for drug-target affinity prediction (DeepDTA), deep learning for compound-protein interaction prediction (DeepCPI), and deep binding affinity refinement (DeepBAR)) are being employed to screen vast chemical libraries rapidly and efficiently. This accelerates the identification of potential lead compounds, saving time and resources; ii) for drug target identification and validation, DL techniques like deep reinforcement learning (DRL) [11] are applied to understand complex biological networks and pathways, aiding in identifying and validating new drug targets with higher precision; and iii) DL models are crucial in absorption, distribution, metabolism, excretion, and toxicity (ADMET) prediction and optimization. They predict the pharmacokinetic and pharmacodynamic profiles of potential drug candidates, thus optimizing drug safety and efficacy. DL models like RNNs and graph neural networks (GNNs) [12] are instrumental in this area, enabling the analysis of complex molecular structures and interactions.

4. NAVIGATING THE COMPLEXITIES AND POSSIBILITIES OF PATIENT STRATIFICATION

From the perspective of patient categorization within biosciences, DL faces significant challenges that also present unique opportunities for advancement. One of the primary challenges is generating ground-truth labels [13] for patient data, which can be expensive or, in some cases, impossible. This is particularly evident in rare diseases or conditions with limited diagnostic expertise. Creating accurate, reliable labels requires substantial clinical expertise and often extensive manual effort, making it a resource-intensive task. Likewise, the inherent variability and complexity of biological data add to the challenge, as ground-truth labels must encapsulate nuanced clinical information accurately.

Data sharing in patient categorization is another critical challenge, primarily due to standardization and privacy considerations. The lack of standardization in medical datasets (e.g. MIMIC-III, Sepsis-3, PhysioNet, ChestX-ray8, Synapse, and i2b2/UTHealth Corpus), including formats and terminologies, complicates the aggregation and comparison of data from dissimilar sources, as expressed in Table 1. This issue is compounded by stringent privacy laws and ethical concerns regarding patient data, which limit data sharing and accessibility. These barriers not only hinder the development of robust DL models (i.e. listed here (e.g. CNNs, RNNs, long short-term memory networks (LSTMs), generative adversarial networks (GANs), deep belief networks (DBNs), autoencoders, transformer models, GNNs, variational autoencoders (VAEs), DRL models, bidirectional encoder representations from transformers (BERT), U-Net, residual networks (ResNets), dense convolutional networks (DenseNets), capsule networks, attention mechanisms, siamese networks, deep convolutional inverse graphics networks (DC-IGN), sequence-to-sequence (Seq2Seq) models, conditional random fields as recurrent neural networks (CRF-RNN)), and explained in the section 5 but also impact their generalizability and applicability across different populations and conditions.

Herewith, Table 1 highlights critical limitations in the datasets used for DL applications in biosciences, which directly impact the accuracy and reliability of model outputs. DL models often rely on large volumes of data, yet the quality and completeness of these datasets can significantly affect the models' learning capacity and generalization abilities. Challenges such as missing data, noise, and inconsistency in annotations can lead to biased learning outcomes, impacting the models' performance across different patient populations. For example, when training on medical datasets like MIMIC-III or PhysioNet, the variation in data collection protocols can introduce inconsistencies that undermine model reliability. Also, representativeness remains a major challenge as these datasets might not fully capture the diversity of real-world patient demographics, leading to models that perform well on specific subgroups but poorly on others. Addressing representativeness is essential for creating models that can generalize effectively to new patient data and mitigate the risks of biased decision-making in clinical practice. Thus, it is evident that specifically in context of Table 1: i) high-quality data ensures that the DL model learns robust patterns rather than fitting to noise, making it crucial for developing predictive models that provide accurate diagnoses and treatments; ii) ensuring representativeness in datasets such as Sepsis-3 or i2b2/UTHealth Corpus prevents the model from favoring specific patient cohorts and ensures that predictions remain valid across a broad population, thereby minimizing biases that could affect clinical decisions; iii) overfitting remains a persistent issue when using complex models on smaller datasets, leading to models that perform well on training data but fail to generalize to unseen data. Mitigating overfitting through techniques like cross-validation and regularization ensures that the model does not memorize patterns specific to a dataset, thus achieving better predictive performance on diverse patient profiles.

Table 1. Comparison of open-access medical datasets for DL: a focus on data quality, size, and ethical considerations

Dataset	Data quality	Data size	Data source and accessibility	Ethical considerations	Regulatory and legal compliance	Long-term sustainability
MIMIC-III [14]	High accuracy, completeness, and timeliness. Moderate representativeness and balance.	Large volume and diversity. Contains data from over 40,000 critical care patients. Limited temporal information.	Collected from a diverse intensive care unit (ICU) population. Freely accessible. Open access with ethical considerations.	De-identified to protect patient privacy, but still carries potential risks of re-identification	Complies with health insurance portability and accountability act (HIPAA) [15] regulations for de-identified data	Supported by MIT Lab for Computational Physiology, ensuring ongoing updates and maintenance. Long-term funding secured through the National Institutes of Health.
Sepsis-3 [16]	High (accuracy, completeness, timeliness).	Moderate (volume, diversity). Specific to sepsis cases	Accumulated from varied hospitals. Unrestricted access with moral considerations. Accessible through request and approval process.	High ethical standards due to the sensitive nature of sepsis data. Potential bias towards specific sepsis phenotypes. Transparent data collection and use.	Adheres to patient privacy laws and sepsis-specific research regulations. HIPAA compliant.	Long-term funding secured.
PhysioNet [17]	High accuracy, completeness, and timeliness. Moderate representativeness and balance.	Large volume and diversity. Extensive temporal information.	Collected from diverse research studies. Access freely available, mindful of ethical aspects.	Limited open access. De-identification and consent protocols in place	Complies with applicable patient data regulations	Supported by a consortium of research institutions, ensuring sustainability.
ChestX-ray8 [18]	High-quality, focused on chest X-ray images.	Large, contains over 100,000 frontal-view X-ray images.	Collected from diverse hospitals. Ethically considerate open access.	Potential bias towards specific chest pathologies. Transparent data collection and use. De-identified, but imaging data can have unique re-identification risks.	Complies with regulations for de-identified imaging data.	Limited long-term funding secured. Likely sustainable due to its utility in AI research and clinical applications
Synapse [19]	Moderate accuracy and completeness, as it is a platform hosting multiple datasets. Moderate representativeness and balance.	Large volume and moderate diversity. Limited temporal information. Accessibility varies by dataset; some are open access while others have restrictions.	Collected from diverse research studies. Openly accessible, maintaining ethical integrity.	Potential bias towards specific brain imaging protocols. Transparent data collection and use. Ethical considerations depend on the specific dataset hosted.	Compliance varies by dataset, generally adheres to standard data sharing regulations.	Long-term financial support obtained, albeit limited. Sustainability depends on the continued support and contribution from the research community
i2b2/UTHealth Corpus [20]	High accuracy and completeness. Moderate representativeness and balance. Limited temporal information.	Moderate, contains clinical narratives and annotations.	Collected from a specific hospital system. Restricted access, requires application and approval for use	Potential bias towards specific patient populations and diagnoses. Transparent data collection and use. De-identified, but text data carries inherent re-identification risks.	HIPAA compliant.	Limited long-term funding secured (i.e. supported by ongoing research collaborations, ensuring its long-term relevance).

5. DEEP LEARNING MODELS

When evaluating DL models, several technical considerations come into play. Firstly, data size and quality are paramount. The amount of data used, data diversity, and data preprocessing techniques all impact model performance. Data cleaning methods and potential biases also need to be addressed to ensure the accuracy and reliability of the results. Data efficiency is another important consideration, as the model's

ability to learn from limited data is crucial in practical applications. The cost, scalability, security, and privacy implications of data acquisition and storage should also be considered [21].

Model architecture, as projected in Figure 1, is another critical aspect to evaluate. Algorithmic details, including the number of layers, neurons, and parameters, impact the model's performance and efficiency. The choice of architecture should be suitable for the specific task at hand and should also consider explainability and interpretability. Hyperparameter optimization, such as the choice and tuning of hyperparameters, plays a significant role in model performance and generalizability. It is essential to find the right balance to ensure optimal outcomes for the DL model.

Table 2 presents a comprehensive overview of the fundamental components of a general DL model adapted for biosciences which is presented in a structured framework. It covers aspects ranging from the architecture of neural networks and identifying structures most suitable for bioscience data to addressing ethical considerations for the responsible handling of sensitive biological information. Each row of the table describes critical elements of the DL modeling process, including data management for training, refinement of model parameters, and ensuring fairness and mitigation of bias. The table is intended as a resource for bioscience professionals, offering insights into leveraging DL methodologies efficiently and ethically, thereby promoting model robustness, interpretability, and adherence to ethical standards.

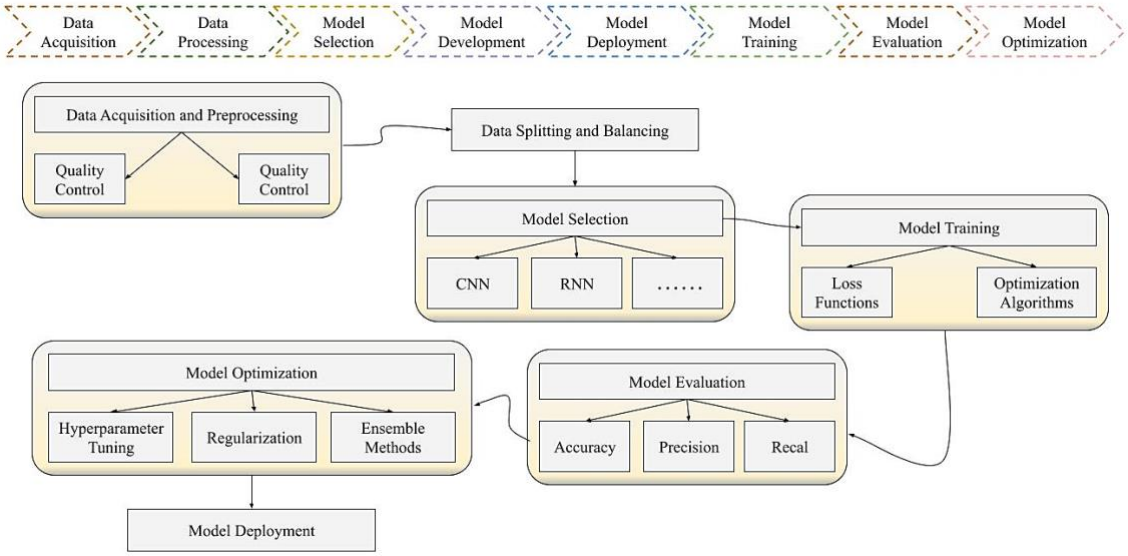

Figure 1. DL model flow for biosciences applications

5.1. Convolutional neural networks

CNNs aim to process data with grid-like topology, exemplified by image data. Their architecture, inspired by the organization of the animal visual cortex, consists of layers designed for tasks like edge detection or pattern recognition. A typical CNN comprises convolutional layers, pooling layers, and fully connected layers [22]. It faces complexities including their 'black-box' nature, sensitivity to hyperparameters, data dependency, and high computational costs. Accurately setting hyperparameters (like filter sizes, and number of layers) is crucial for model performance. Their data dependency is evident in performance which varies with data quality and quantity. CNNs, especially with large-scale data, require substantial computational resources like GPUs or tensor processing unit TPU and efficient algorithms. Their accuracy and efficiency are paramount, alongside generalizability and robustness against variations in input data. Interpreting and explaining CNN decisions remains a challenging due to high-dimensional representations and complex decision boundaries. This complexity often obscures the causal relationships within the model.

Thus, CNNs must address ethical considerations and regulatory compliance, particularly regarding bias, fairness, transparency, and accountability [23]. Ensuring that CNNs do not perpetuate existing biases and are accountable in their decision-making processes is crucial in their application, particularly in sensitive fields like healthcare systems. The taxonomy of DL models, as illustrated in Figure 2, provides a structured overview of various architectures and their respective sub-types, categorizing them into distinct groups such as convolutional, recurrent, generative, and more. This classification aids in understanding the diversity and specialization of models (elaborated from subsection 5.1 to 5.20) in the field, highlighting their unique bioscience applications and evolutionary paths in AI research.

Table 2. DL modelling considerations for biosciences

Category	Parameter and optimality criteria	Considerations and hardware requirements	Data preprocessing and model robustness	Ethical and reproducibility concerns
Architecture	Design of neural network layers	Bio-specific constraints, GPU/CPU needs	Preprocessing genomic/imaging data	Ethical use of biological data
Hyperparameters	Tuning for optimal performance	Limited compute resources handling	Data normalization, augmentation	Consistency in varied bioscience data
Training data	Volume, variety, veracity	Storage and processing capacity	Handling imbalanced datasets	Patient data privacy
Evaluation metrics	Accuracy, sensitivity, specificity	Requirement alignment for biosciences	Robustness to noisy/uncertain data	Impact on clinical decisions
Evaluation tracking management	Tools for model tracking	Infrastructure for large-scale studies	Stability across different datasets	Transparency in clinical trials
Model compression and pruning	Efficient model size reduction	Resource constraints in bio labs	Preserving essential features	--
Knowledge distillation	Transfer learning efficiency	Hardware for complex models	Maintaining biological relevance	--
Learning matrix	Adaptability to bio datasets	Specialized hardware for learning	Feature extraction from complex data	--
Interpretability	Model decision rationale	--	--	Clarity in clinical decision support
Fairness and bias Mitigation	Equal performance across groups	--	Balancing datasets	Avoiding bias in patient treatment
Piracy and security	Data protection mechanisms	Secure storage and transfer	--	Patient confidentiality
Deployment criteria	Real-world usability standards	Infrastructure in clinical settings	Validation on real patient data	Compliance with healthcare standards
Uncertainty quantification	Handling probabilistic outcomes	--	Statistical methods for uncertainty	Reliability in diagnostic tools
Explainability	Clarity in model outcomes	--	--	Justifiability in treatment decisions

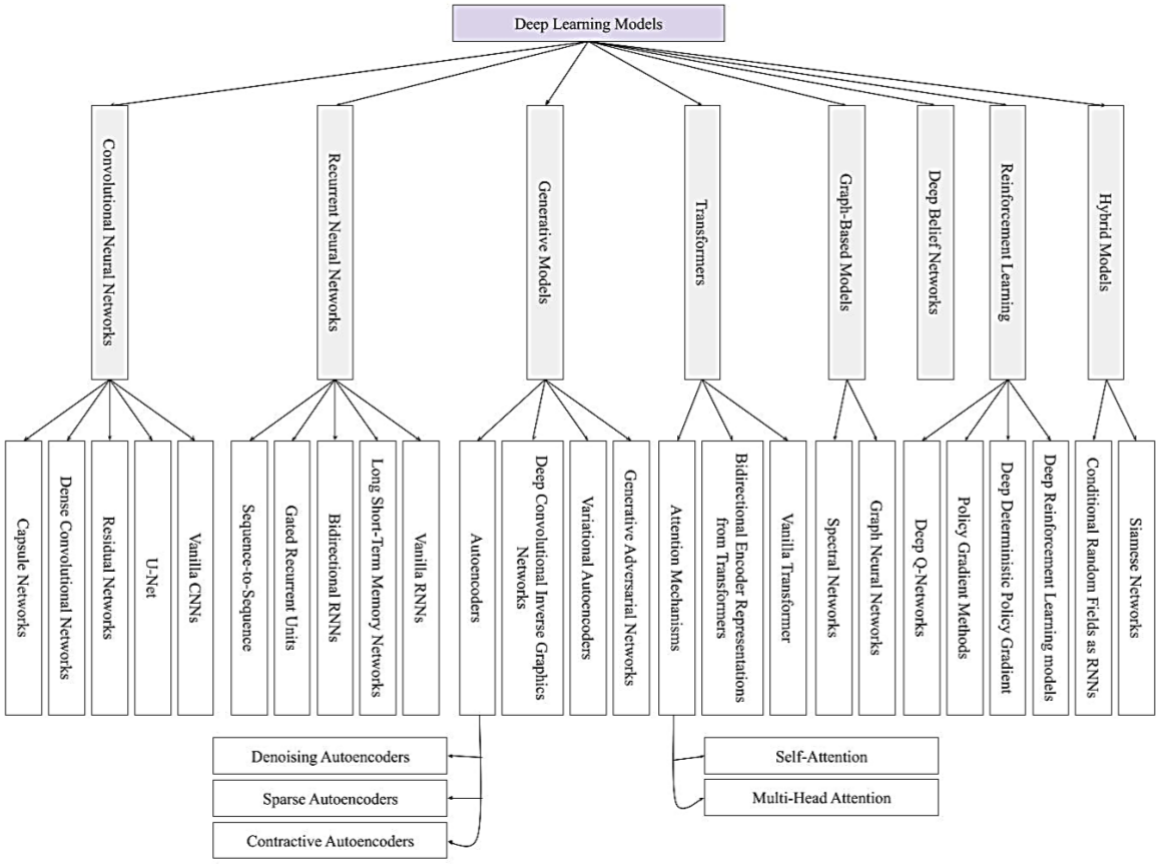

Figure 2. Comprehensive taxonomy of DL models applicable in biosciences

5.2. Recurrent neural networks

RNN excels at processing sequences, making it invaluable in the biosciences. The model tackles tasks like protein structure prediction, gene expression analysis, and drug discovery [24]. This capability stems from its "hidden state," which carries information across time steps, enabling the pattern to learn complex relationships within sequences. Nonetheless, implementing RNNs presents significant challenges. For instance, dataset selection de-pends on the intended application. In the context of Table 1, MIMIC-III, a critical care database, fuels RNNs to predict patient outcomes and personalize treatment. Sepsis-3, containing clinical data for sepsis patients, helps detect sepsis early. PhysioNet, a repository of physiological data, empowers RNNs to analyse signals and identify abnormalities. ChestX-ray8, with labelled chest X-ray images, allows automatic disease detection. Synapse, an open-source platform, grants access to medical imaging datasets for tasks like image segmentation and disease progression prediction. The i2b2/UTHealth Corpus, containing de-identified clinical narratives, enables RNNs to extract valuable clinical information from text.

Although RNNs hold significant promise, they encounter several obstacles. The opaque nature of RNNs complicates interpretability which leads to challenges in trusting their outputs. Furthermore, they exhibit high sensitivity to hyperparameter settings and need substantial datasets for effective performance, which can be problematic in biosciences where data is often scarce. Training RNNs also demands considerable computational power, often requiring access to high-performance resources, making their implementation both complex and resource-intensive.

5.3. Long short-term memory networks

LSTMs are a type of RNN that excels at capturing long-range dependencies in sequential data (i.e. data related to deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and amino acids). The main objective of LSTMs is to process and predict sequences of data by learning from patterns in the past. LSTMs require sequential data, such as time series or natural language, where the order of the data points matters. Our investigation revealed that the model holds following characteristics [25]:

- LSTMs are powerful deep-learning models for processing sequential data.
- It comes with complexities related to their black-box nature (such as long-term dependency capture, gated cell architecture, hidden state propagation, gradient vanishing/exploding mitigation, and parameter efficiency), sensitivity to hyperparameters (i.e. learning rate, batch size, number of hidden units, regularization techniques, gradient clipping, sequence length, loss function, and data processing), data dependency (i.e. data noise, missing data, temporal dependence, order dependence, stationarity, and data distribution), and computational cost.
- Evaluating the model's accuracy, efficiency, generalizability, and robustness is fundamental.
- Interpreting and explaining LSTMs can be challenging due to high-dimensional representations, complex decision boundaries, and the lack of causal explanations.
- Computational resource constraints and ethical considerations, such as bias and fairness, transparency, and accountability, need to be carefully addressed when utilizing LSTMs.

Herewith, the LSTMs aim for high precision in sequence prediction while maintaining computational efficiency. The ability of the model to generalize well to new and unseen data is equally essential for robust performance. Thus, interpreting and explaining LSTMs can be challenging due to their high-dimensional representations, making it difficult to visualize and understand the inner workings of the model. Complex decision boundaries further complicate the interpretability, as LSTMs can capture intricate patterns that are not easily explained in simple terms. Due to the intrinsic nature of the model, LSTMs lack causal explanations, meaning that they can predict outcomes based on patterns without providing a clear understanding of the underlying cause-effect relationships.

5.4. Generative adversarial networks

GNNs can generate new data with the same statistics as the training set. They are commonly used in the field of biosciences to learn the generative model of any data distribution through adversarial methods (such as data poisoning, reconstructing sensitive patient information from GAN-generated medical images, and manipulating interpretability results to hide malicious intent or bias) with the help of deep neural networks [26]. GANs can be used to generate new samples of biological systems (i.e. DNA and RNA sequences, organoids, microbiomes, and biological imaging), which were analyzed by our research collaborators to understand the system better. GANs can also be used to prepare samples for study, direct research, and model biological systems. While we tested the model effectiveness, we benefited from the following characteristics:

- Capable of realistic generation of synthetic data for tasks like protein structure prediction, drug discovery, and medical imaging.
- Eligible for designing novel molecules with desired properties for drug development and materials science.
- Effective in enhancing low-resolution biological images to improve diagnosis and analysis.

- Impactful in discovering hidden patterns and relationships in large biological datasets without needing labeled data.
- Efficient in creating new features from existing data to improve model performance and interpretability.
- Applicable in predicting protein structures from amino acid sequences, aiding functional understanding and drug design.

5.5. Deep belief networks

The primary objective of DBNs is to extract rich internal representations of complex biological data, such as gene expression patterns, protein interactions, and drug discovery. DBNs achieve this by combining unsupervised learning principles and neural networks, allowing them to learn a generative model of the input data. This enables them to probabilistically reconstruct their inputs, making them particularly suitable for tasks such as pattern recognition, feature extraction, and data modeling in bioinformatics and computational biology.

DBNs are trained on several types of biological data, including gene expression data, protein sequences, and chemical compound structures. They are used to identify patterns in the data that can lead to the development of novel diagnostic tools, drug discovery, and predictive modelling in biosciences [27]. Despite their effectiveness in learning complex representations from biological data, it is important to note that DBNs have fallen out of favor in some areas of machine learning, and their usage may be limited in certain contexts.

Unlike traditional feedforward DNNs, the DBNs are composed of numerous layers of latent variables with associations between the layers but not within every layer. They are trained layer-by-layer using unsupervised learning algorithms, such as restricted Boltzmann machines (RBMs), to learn a generative model of the input data. This unsupervised pretraining allows DBNs to learn a stable representation of the data, which can be further fine-tuned with supervised learning for tasks such as classification or regression.

5.6. Autoencoders

Autoencoders are a class of neural networks used in unsupervised learning, aiming to capture efficient representations of the input data. In the context of biosciences, autoencoders can be employed for tasks such as feature extraction from high-dimensional biological data, denoising of gene expression profiles, or dimensionality reduction for single-cell RNA sequencing data. The primary objective of an autoencoder is to minimize the reconstruction error, which can be formulated as the mean squared error or binary cross-entropy loss. They can handle various data types, including continuous, categorical, and binary data, making them suitable for diverse biological datasets. A few observed limitations of using autoencoders in biosciences include:

- The need for a large and representative training dataset, as they can deliver mixed results if the dataset is not large enough, clean, or too noisy.
- It can be sensitive to input errors and may eliminate important information in the input data, leading to data loss.
- The black-box nature of autoencoders and difficulty in interpreting the learned representations, especially in the context of high-dimensional biological data, pose significant limitations in their application in biosciences.
- The potential for the latent space learned by autoencoders to be discontinuous, making it difficult for easy interpolation and random sampling, which can limit their utility in certain applications (such as feature extraction, dimensionality reduction, image denoising, and anomaly detection in images) enabling the identification of unusual or irregular patterns that may indicate a different class or category.
- The specific training of autoencoders to learn and reproduce input features, which generates algorithms that may not work as well for new data, can be a constraint in biosciences, where generalizability and robustness are crucial.

5.7. Transformer models

Transformer models have been increasingly used in various applications, including protein property prediction, drug discovery, and analysis of genetic data. These models have exhibited promising results due to their ability to learn long-range reliance and complex models in the data, making them particularly appropriate for sequence-based problems. Some technical functions and model specifications of transformer models in biosciences, includings:

- Transformer models use multi-head attention mechanisms to capture multiple features simultaneously, allowing them to learn complex patterns and representations in the data.
- These models also incorporate fully connected feed-forward networks as intermediate components, which help to learn and capture long-range dependencies in the input data.
- Transformer models in biosciences are often trained on large datasets, qualifying them to learn complex patterns and representations from the data.
- Transformer models have been used to predict protein properties, such as stability, solubility, and post-translational modifications, by learning complex patterns in the protein sequence and structure.

– Transformer models have been adopted for protein-specific de novo drug generation, treating the problem as a machine translation task.

Advantages of transformer models in biosciences include their ability to learn long-range dependencies, capture complex patterns in the data, and handle large-scale training, which allows them to learn meaningful representations from large datasets. Nonetheless, some limitations of Transformer models in biosciences include their high computational cost, difficulty in interpretability, and potential for overfitting. Which may require careful regularization techniques and additional validation steps.

5.8. Graph neural networks

GNNs are transforming biological data analysis by modelling complex networks, such as protein-protein interactions and gene regulatory networks, using graph representations like adjacency matrices and edge lists. These networks utilize message-passing mechanisms like sum, average, or gated to learn node representations, a key aspect in understanding biological functions and relationships. GNN architectures, including graph convolutional networks (GCN), GraphSage, and graph attention network (GAT) [28] are adept at learning node embeddings, capturing the nodes' intrinsic properties and network interactions. While their scalability and interpretability are advantageous for handling diverse, large biological networks, they face challenges related to data quality, model explainability, and computational costs. Nevertheless, GNNs have shown remarkable applications in biosciences, from predicting protein-protein interactions to assisting in drug discovery and biomarker identification, demonstrating their potential to significantly advance biosciences research despite some limitations, such as:

– Its performance profoundly depends on the quality and completeness of the input data, particularly in biological networks where data can be sparse or incomplete.
– Despite their competence, GNNs are often considered 'black-box' models, making it difficult to interpret how they arrive at certain conclusions, a critical aspect in biosciences for validating findings.
– Training GNNs, especially on large biological networks, demands significant computational power and resources, which can be a limiting factor.
– While GNNs are scalable, handling extremely large and complex biological networks can still be challenging, affecting their efficiency and performance.
– Incorporating existing biological knowledge into GNN models remains a challenge, limiting their effectiveness in certain applications (such as, disease pathogenesis analysis, drug repurposing, complex trait prediction, metabolic pathway analysis, cellular development studies, and evolutionary biology research).

5.9. Variational autoencoders

VAEs unlock the hidden language of life, extracting patterns from complex data and generating novel entities for discovery. They differ from traditional autoencoders by their ability to generate new, plausible data points by learning a distribution over the input data. VAEs consist of two independent components: an encoder that records input data to a hidden space, and a decoder that recreates facts from this space [29]. A key aspect is the Kullback-Leibler divergence [30], which regularizes the latent space to improve generalization.

In the perspective domain specific to biosciences, VAEs are advantageous for their capacity to handle high-dimensional data like genomic sequences and protein structures, providing interpretable representations. They are instrumental in generating novel biological data, aiding in areas like drug discovery and genomics. However, they face limitations such as high computational demands, sensitivity to hyperparameters, and complexities in interpreting the latent space. Conclusively, the VAEs power a growing toolbox for biologists, generating synthetic data, predicting molecular properties for drug discovery, and guiding protein engineering through structural modelling. As researchers refine their architecture and enhance their interpretability, VAEs promise to become even more efficient and accessible, unlocking new frontiers in complex biological data analysis.

5.10. Deep reinforcement learning models

DRL models use deep neural networks as function approximators to handle high-dimensional input spaces and complex environments more effectively than traditional reinforcement learning (RL) methods. The core architecture of the model consists of an agent, which interacts with an environment, receives feedback through a reward function, and makes decisions based on a policy network [31]. The learning paradigm driving the model's behavior can be q-learning, which involves estimating the value of an action in a particular state, or policy gradient, which directly optimizes the policy network to maximize the expected cumulative reward. During training, the DRL algorithm typically involves exploration vs. exploitation strategies, where the agent balances between trying out new actions and exploiting the best-known actions, experience replay to break correlations between consecutive state-action pairs, and gradient updates for the policy network to improve decision-making over time.

One of the key strengths of DRL models is that they can learn from trial-and-error interactions, enabling them to adapt to diverse data sources and make decisions in situations where the outcomes are not immediately clear. However, DRL models have limitations, including their computational cost, data requirements, potential for convergence issues, and challenges in interpreting the learned policy. Despite these limitations, the potential of DRL in biosciences is significant, and our collaborative ongoing research is exploring promising future applications, such as model-based DRL for multi-agent pathfinding in complex and crowded environments.

5.11. Bidirectional encoder representations from transformers

The BERT's architecture is based on the transformer model, which allows it to capture context from both the left and right side of a token's position within the text. This is achieved through attention mechanisms that weigh the importance of different tokens in understanding the meaning of a given word [32]. In biosciences, BERT's ability to understand the context and relationships within biological texts offers several advantages: i) enhanced understanding of complex biological terminologies and interactions; ii) improved performance in tasks like biomedical text mining, genomic data analysis, and protein structure prediction. Conversely, applying BERT comes with challenges, for example: i) The need for large, annotated datasets specific to biosciences; ii) High computational resources are required for training and inference. Based on BERT's usability pros and cons, the applications in biosciences include: i) Analysing genomic sequences and predicting protein structures; ii) Extracting information from biomedical literature and clinical notes. Compared to other DL models, BERT's effectiveness in understanding context and handling natural language tasks stands out. Yet, its computational requirements (such as, large memory requirements for processing extensive models and datasets, and dependence on large, well-annotated biological datasets for effective training) and the need for extensive training data can be limiting factors in certain bioscience applications.

5.12. U-Net

U-Net is CNN exclusively conceived for biomedical image segmentation, with a u-shaped architecture consisting of a contracting path and an expansive path. The network is based on a fully CNN and can segment a 512×512 image in less than a second on a modern GPU. U-Net has been widely used for various quantification tasks in biosciences, such as cell counting, detection, and morphometry [33].

The underlying algorithm of U-Net is based on DL, allowing the network to classify each pixel in the input image, enabling precise segmentation and localization of borders. Its unique features include the use of transposed convolutional layers in the expansive path, which enables it to upsample and refine the segmentation results. U-Net's strengths lie in its ability to handle complex tasks, making it suitable for a wide range of domain specific application.

In specified context (i.e. biosciences), the advantages of the U-Net model include its excellent performance in biomedical image segmentation tasks, outperforming traditional methods such as sliding-window convolutional networks. It has also been applied to various image modalities, such as fluorescent stains and transmission electron microscopy (TEM) [34]. Conversely, U-Net's performance may vary depending on the specific dataset and experimental conditions, requiring fine-tuning and optimization for each new application. Accordingly, the network's architecture may not be suitable for all tasks, and alternative models may be more appropriate in certain cases.

Thus, U-Net has found wide-ranging applications in biosciences, including cell counting and detection, morphometry and shape measurements, medical image reconstruction, and image-to-image translation for estimating fluorescent stains. These applications have significantly contributed to the advancement of biomedical research, enabling more effective analysis, and understanding of cellular processes, disease diagnosis, and drug discovery.

5.13. Residual networks

ResNets address the vanishing gradient challenge in deep neural networks by instituting skip connections, allowing gradients to flow through a network without hindrance. This enables the construction of much deeper networks, with layers simply learning residual functions with orientation to the layer inputs, rather than learning unreferenced functions [35]. The model offers significant advantages due to its depth and capacity for feature extraction. The depth of a ResNet, characterized by numerous layers and parameters, enhances its ability to learn complex patterns in biological data, which is often high-dimensional and intricate. This depth, however, introduces challenges in terms of model complexity and computational requirements. Training such deep networks requires substantial computational resources and careful hyperparameter tuning to optimize performance and prevent overfitting.

ResNets have found applications in various bioscientific domains, like image-based diagnosis, genomic data analysis, and protein structure prediction. Their ability to effectively handle large datasets and extract meaningful patterns from complex biological data makes them a valuable tool in these areas.

However, the complexity and resource requirements of ResNets, alongside challenges in model interpretability and data dependency, are important considerations in their application in biosciences.

5.14. Dense convolutional networks

DenseNets key feature lies in dense connections between layers, fostering efficient feature reuse and alleviating the vanishing gradient problem. It boasts a unique architecture that redefines feature reuse and information flow within CNN. Unlike traditional models where each layer only receives input from the preceding one, DenseNets connects every layer to every other layer in the network, fostering a dense web of information exchange [36]. Imagine a pyramid where each layer is not just stacked on top of the one below, but intricately woven with data threads flowing back and forth. This dense connectivity grants DenseNets several advantages, such as reduced vanishing gradients, improved feature propagation, and parameter efficiency. Along with the pros, we have noticed a few limitations, such as increased memory footprint, hyperparameter tuning complexity, and limited explainability.

Despite these challenges, DenseNets remains a powerful tool for various tasks in computer vision, bioinformatics, and medical imaging. Their ability to handle complex data efficiently, achieve high performance, and learn from limited datasets makes them well-suited for tasks like protein structure prediction, medical image analysis, and disease diagnosis. As research progresses in areas like efficient compression techniques and improved interpretability methods, DenseNets are prepared to play an even greater role in unlocking the secrets of biology and improving healthcare outcomes.

5.15. Capsule networks

Capsule networks (CapsNets) introduce a novel architecture in DL, addressing the limitations of traditional neural networks, especially in handling spatial hierarchies in data. It consists of capsules, and groups of neurons that encode both the probability of an entity's presence and its spatial orientation, a significant shift from the scalar outputs of conventional neurons. This structure enables CapsNets to understand the spatial relationships in data, which is crucial in complex bioscience imaging tasks [37].

It can effectively work with smaller datasets, a common scenario in specialized bioscience research, and is particularly adept at recognizing patterns and features from diverse and complex data types. Nevertheless, they require meticulous data preprocessing and are sensitive to data quality and biases. It also faces challenges in scalability and data storage, given its complex architecture. For instance, CapsNets are more intricate than traditional CNNs, with multiple layers and dynamic routing algorithms, making them computationally intensive. This complexity necessitates robust hardware (i.e. with features, such as high-performance GPUs, large memory capacity, fast processors, high-bandwidth storage, advanced cooling system, and scalable architecture) for training and optimization. Despite these challenges, CapsNets hold promise in biosciences for tasks like protein structure prediction and biomedical image analysis, offering enhanced accuracy and efficiency.

5.16. Attention mechanisms

Attention mechanisms allow bioscience-focused models to concentrate on specific parts of the data, improving efficiency and accuracy, especially in handling diverse and complex biological datasets [38]. They are particularly adept in limited data, enhancing the model's ability to learn from small datasets. The architectural complexity of attention-based models, such as the number of layers and parameters, requires careful hyperparameter optimization to ensure optimal performance. This complexity can increase computational costs and energy consumption during training and inference, necessitating robust hardware.

These models are crucial in tasks like genomic sequence analysis, protein structure prediction, and biomedical image processing. Their ability to provide insight into model decisions adds transparency, aiding ethical considerations like bias detection. However, challenges in scalability, data privacy, and security remain key considerations in deploying these models in bioscience applications.

5.17. Siamese networks

Siamese networks [39], featuring twin structures connected at the output, excel in tasks like analysing genetic sequence similarities. Efficient even with limited data, they require careful data preprocessing to avoid bias. Their mirrored architecture necessitates precise hyperparameter adjustments for optimal performance, aligning well with bioscience tasks needing relational analysis. Nonetheless, they pose challenges in computational intensity and model interpretability, especially in complex bioscience applications. Requiring robust computing resources, they also demand attention to ethical issues like fairness and transparency. Despite these hurdles, their potential in areas like comparative genomics and personalized medicine, harnessing their pattern comparison capabilities is noteworthy.

5.18. Deep convolutional inverse graphics networks

DC-IGN model [40] was designed to learn the essential structure of an image and generate a 3D model of the object depicted in the image. The model is instructed on a large dataset of images and uses a

CNN to extract features from the images. The extracted features are then used to generate a 3D model of the object. The architecture consists of several layers, each responsible for obtaining explicit attributes from the input data. The model typically includes an encoder-decoder structure, with the encoder responsible for feature extraction and the decoder for generating the final output. The specifications, such as the number of layers, neurons, and parameters, can vary depending on the specific task and dataset. Then again, a common architecture might include several convolutional layers, followed by a series of completely coupled layers, and finally an output layer for generating the final output.

This model differs from traditional CNNs in its primary focus and attention. Traditional CNNs are mainly used for supervised assignments like classification, without a specific emphasis on straighten out the factors of alteration extant in the input data. In contrast, DC-IGNs are designed to reveal the primary structure of the data and produce images with explicit properties, such as outline, model, and illumination conditions. This is achieved by using an unsupervised learning model that entails of both an analysis network and a synthesis network, allowing the model to learn a mapping from bioscience-focused images to a set of latent variables that represent the underlying structure of the data. This architecture also has several potential drawbacks over other CNNs. Some of these drawbacks include:

- It relies on large datasets to learn meaningful representations of images. If the dataset is limited or of poor quality, the model may struggle to learn effective representations, impacting its performance on real-world tasks.
- It often involves deep CNN, which can be computationally steep and resource exhaustive. This can limit the applicability of DC-IGNs in situations where computational resources are limited, or cost is a concern.
- It can be challenging to train due to the deep network architecture and the complexity of learning disentangled representations of images. This can direct to longer training times and potentially more computational resources than traditional CNNs.

5.19. Sequence-to-sequence models

Seq2Seq models typically consist of an encoder-decoder architecture with RNN such as LSTM or gated recurrent unit (GRU) networks. The encoder-decoder architecture is designed to convert input sequences into a fixed-size representation, which is then used to generate the output sequence [41]. In general, the architecture of Seq2Seq models includes: i) encoder, that is responsible for processing the input sequence. It is typically implemented using an RNN like LSTM or GRU, which can handle long-term dependencies in sequences; ii) decoder, that generates the output sequence based on the fixed-size representation obtained from the encoder. It can also use attention mechanisms to improve performance for more complex tasks. Due to models' specific architecture, it faces several challenges, including: i) models can struggle with long sequences because they rely on recurrent layers or self-attention mechanisms; ii) it can sometimes exhibit instability (such as vanishing gradients, mode collapse, overfitting and underfitting, and training instability), leading to poor performance or the need for extensive fine-tuning.

5.20. Conditional random fields as recurrent neural networks

CRF-RNN combines the strengths of CRFs and RNNs for sequence modeling. This hybrid model [42] leverages the sequential data processing capabilities of RNNs and the structured output prediction strengths of CRFs. It is particularly suited for tasks where context and sequence relationships are critical.

The CRF layer typically sits atop the RNN, enabling the model to make contextually informed predictions. This setup is beneficial for complex sequence modeling tasks, commonly found in bioscience applications like genomics and protein structure analysis. The combination aims to harness both the DL power of RNNs and the precision of CRF's output structuring.

Looking forward, the future vision for Seq2Seq models like CRF-RNN includes advancements in computational efficiency, enhanced model interpretability, and resolving ethical challenges. As per described in sections 5.1-5.20, it is evident that in the training and optimization stage, the choice of the training algorithm, optimizer, and learning rate directly affects the model's convergence speed, stability, and performance. Regularization techniques, such as L1 and L2 regularization and dropout, help prevent overfitting and improve generalizability. It is also important to consider the hardware and software used for training, as their performance, cost, scalability, and energy consumption can significantly impact the overall training process.

Regarding inference and deployment, the model's size and latency are crucial factors. The size of the model after training affects inference speed and deployment cost. The memory footprint of the model during inference should be considered to ensure its suitability for different hardware platforms. Methods for model serving and scaling in production environments, along with their associated costs and latency, need to be carefully evaluated. Nevertheless, the performance and evaluation metrics are vital for assessing the model's effectiveness. Choosing appropriate metrics for specific tasks, such as accuracy, precision, recall, and F1-score, ensures accurate evaluation. Generalizability, or how well the model performs on unseen data and its

transferability to different tasks or domains, is another important consideration. Explainability and interpretability, which involve understanding how the model makes decisions and the importance of features [42], contribute to reliable evaluations.

6. QUANTUM COMPUTING AND DEEP LEARNING IN BIOSCIENCES

Quantum neural networks (QNNs) [43] in biosciences are focused on modelling complex biological systems. They utilize quantum principles to efficiently process large biological datasets. In drug discovery, quantum computing is being explored for its potential to analyse molecular interactions at the quantum level, which could significantly enhance the accuracy of predictions for drug candidates. The main challenges include developing robust quantum hardware and algorithms tailored for biological complexities. As quantum computing technology matures, it's anticipated to significantly advance drug discovery processes and provide deeper insights into biological systems.

6.1. Artificial intelligence and pandemic response

Quantum computing and AI are pivotal in pandemic response, enhancing disease spread prediction and mutation analysis [44]. DL models, leveraging quantum computational power, enable rapid, large-scale analysis of pathogen genomes, predicting mutation patterns and spread dynamics with greater accuracy. This quantum-AI synergy is crucial for developing vaccines, where AI models assist in identifying potential vaccine candidates and optimizing their efficacy. In epidemic modelling, AI algorithms, backed by quantum computing, offer nuanced insights into disease dynamics, facilitating more effective containment strategies. These advancements highlight the potential of quantum-enhanced DL in managing public health crises.

6.2. Quantum computing steered deep learning-focused agricultural biosciences

In agricultural biosciences, integrating quantum computing with DL is revolutionizing approaches to crop disease prediction and soil analysis. AI models, enhanced by quantum computational capabilities, are increasingly being used to accurately predict disease outbreaks in crops, enabling pre-emptive measures to protect yields. These models analyse complex patterns in large datasets, including climatic factors and plant genetics, to identify potential threats. Similarly, machine learning techniques are instrumental in soil analysis, providing detailed insights into soil health and nutrient profiles, which are vital for sustainable farming practices. This technological synergy also plays a crucial role in enhancing food security, enabling the development of more efficient and targeted agricultural strategies, optimizing resource use and crop yields. Applying quantum-enhanced AI in agriculture underscores its potential in addressing global food security challenges.

6.3. Unraveling quantum computing focused deep learning aided neuroinformatic and neurodegenerative disease research

It is evident that the fusion of quantum computing and DL in neuroinformatics is reshaping our understanding of neural networks and brain function. Quantum-enhanced AI models (e.g. quantum convolutional neural networks (QCNNs) [45], quantum recurrent neural networks (QRNNs) [46], quantum Boltzmann machines (QBMs) [47], quantum support vector machines (QSVMs) [48], quantum graph neural networks (QGNNs), quantum reinforcement learning (QRL) models, quantum variational autoencoders (QVAEs), quantum principal component analysis models (QPCA) [49], quantum decision trees [50], and quantum generative adversarial networks (QGANs)) offer unprecedented computational power to analyze vast neural datasets, facilitating deeper insights into the complexities of brain function, neural network behavior, user activity behavior [51]. This approach is particularly beneficial in researching neurodegenerative diseases (such as Alzheimer, Parkinson, and Huntington), where DL algorithms can detect subtle patterns and changes in neural data, potentially leading to early diagnosis and more effective treatments.

6.4. Impact on environmental biosciences in ecosystem modelling

Environmental bioscience is undergoing a revolutionary transformation, fuelled by the synergistic power of quantum computing and DL. Thus, advanced AI models (e.g. quantum ecosystems network (QEN), variational autoencoders for biodiversity mapping (VAE-Bio) [51], and ecosystem evolution simulator (EES)) that are empowered by quantum processing muscle, are tackling the intricate complexities of ecological data, unveiling with unprecedented precision the hidden dances of ecosystem dynamics and the delicate tapestry of biodiversity. This newfound clarity paves the way for the development of targeted and effective conservation strategies, safeguarding the very fabric of our planet. DL, meanwhile, stands as a sentinel in the face of climate change. Its keen eye scrutinizes vast datasets, deciphering the whispers of long-term climatic trends and their impact on ecosystems. This crucial intelligence informs future climate scenarios, guides risk assessments, and empowers the development of data-driven mitigation strategies, marking a decisive shift toward an era of proactive environmental management. Hence, quantum and DL, hand in hand, are shaping a future where environmental understanding and proactive action converge, safeguarding the precious balance of our planet.

7. ETHICAL AND REGULATORY CONSIDERATIONS IN DEEP LEARNING FOR BIOSCIENCES

DL ethics in biosciences involve critical issues like ensuring data privacy, avoiding bias in AI models, and ensuring transparency in AI-driven decisions. These are crucial in maintaining public trust and adhering to ethical research practices. Ethical considerations in DL-driven biological research extend to responsible data handling, informed consent for data use, and equitable access to DL-driven technologies. Regulatory frameworks are needed to oversee DL applications in sensitive areas like genetic research and patient data analysis, ensuring they conform to ethical standards (such as, the Belmont Report [52], the declaration of Helsinki [53], and the asilomar AI Principles [54]). Responsible DL in biomedical data maintains a balance between advancing innovation and addressing privacy and ethical issues. Approaches involve:

- Implementing robust data privacy measures.
- Ensuring fairness and avoiding biases in DL models [55], [56].
- Transparency in DL algorithms and decision-making processes [57].
- Adhering to regulatory standards and ethical guidelines.
- Regular auditing of DL systems for ethical compliance.
- Developing DL models with interpretability to allow users to understand how decisions are made.
- Ensuring inclusivity and diversity in training datasets to prevent skewed DL outcomes.
- Collaborating with ethicists and biologists during DL system development.
- Fostering open and transparent communication about DL technology's capabilities and limitations.
- Developing contingency plans for DL failures or unintended consequences.
- Encouraging multi-disciplinary collaboration to address complex ethical issues.
- Aligning DL development with societal values and healthcare norms.
- Engaging in ongoing dialogue with regulatory bodies to shape evolving DL governance frameworks.

8. PROSPECTIVE TRENDS IN DL IMPLEMENTATION WITHIN BIOSCIENCES

We envision that the future directions of DL in biosciences are poised to leverage advancements in computational power and algorithmic efficiency. A critical focus will be on integrating multi-omic data sources [55] using sophisticated neural network architectures. This integration will require leveraging cloud computing platforms and distributed computing frameworks like Apache Spark, enabling real-time analysis of vast datasets. Enhanced GPU acceleration, possibly through next-generation NVIDIA CUDA cores [58] or Google's TPU technology, will drive these analyses, reducing computational time for tasks like genomic sequencing analysis or complex molecular simulations.

Alternative promising direction is the application of DRL in drug discovery and protein folding problems. Advanced DRL algorithms, coupled with high-throughput screening methods, could significantly expedite the identification of potential drug candidates. Besides, the integration of quantum computing into DL frameworks, using tools like TensorFlow quantum, may revolutionize our approach to computational problems in structural biology, offering exponential speedups in simulations of molecular dynamics.

The use of federated learning in healthcare data will also gain prominence. This approach, which allows for decentralized data processing while maintaining privacy, is crucial in handling sensitive patient data. Tools like PySyft [59] or TensorFlow Federated will become essential in creating robust models that learn from distributed data sources without compromising individual data privacy. These technologies will be instrumental in advancing personalized medicine, enabling researchers to develop more accurate predictive models for disease progression and treatment outcomes.

Looking ahead, potential breakthroughs in DL applications within biosciences could involve the development of self-learning models capable of continuously adapting to new data and evolving without human intervention. Such models would incorporate mechanisms like meta-learning and continual learning to autonomously update their parameters based on real-time data streams, allowing for dynamic adaptation in clinical settings where data characteristics change rapidly. Another promising direction lies in the fusion of neural-symbolic reasoning, which combines DL with symbolic AI techniques to create systems that not only recognize patterns but also understand and reason about biological processes, potentially enhancing the interpretability of complex models used in genomic analysis and personalized medicine. Furthermore, advancements in neuromorphic computing, inspired by the architecture of the human brain, could enable the design of highly energy-efficient DL systems, making it feasible to deploy sophisticated models directly at the edge for real-time biosignal processing, which is crucial for applications such as wearable health monitoring and mobile diagnostic devices. These potential advancements may offer a pathway toward achieving unprecedented levels of accuracy, interpretability, and scalability in DL solutions tailored to the unique challenges of biosciences.

9. CONCLUSION

Consequently, the exploration of DL in biosciences has unveiled a myriad of possibilities and challenges. The convergence of advanced computational models with complex biological systems has set the stage for groundbreaking discoveries in drug development, genomics, and personalized medicine. Future advancements hinge on the synergistic integration of emerging technologies such as quantum computing, federated learning, and multi-omic data analysis, promising to reshape our understanding of biological processes. Nonetheless, these advancements must be navigated with ethical considerations and a focus on data privacy. This dynamic field stands at the cusp of a new era, where the amalgamation of AI and bioscience could lead to unprecedented breakthroughs, offering profound implications for healthcare and beyond.

ACKNOWLEDGEMENTS

This study was sponsored by Prince Sattam bin Abdulaziz University through project number 2023/RV/8.

REFERENCES

[1] M. Pandey *et al.*, "The transformational role of GPU computing and deep learning in drug discovery," *Nature Machine Intelligence*, vol. 4, no. 3, pp. 211–221, Mar. 2022, doi: 10.1038/s42256-022-00463-x.
[2] A. Chattopadhyay and M. Maitra, "MRI-based brain tumour image detection using CNN based deep learning method," *Neuroscience Informatics*, vol. 2, no. 4, Dec. 2022, doi: 10.1016/j.neuri.2022.100060.
[3] F. Pouromran, Y. Lin, and S. Kamarthi, "Personalized deep Bi-LSTM RNN based model for pain intensity classification using EDA Signal," *Sensors*, vol. 22, no. 21, Oct. 2022, doi: 10.3390/s22218087.
[4] B. Dou *et al.*, "Machine learning methods for small data challenges in molecular science," *Chemical Reviews*, vol. 123, no. 13, pp. 8736–8780, Jul. 2023, doi: 10.1021/acs.chemrev.3c00189.
[5] R. Tang *et al.*, "Low-latency label-free image-activated cell sorting using fast deep learning and AI inferencing," *SSRN Electronic Journal*, 2022, doi: 10.2139/ssrn.4177986.
[6] N.-C. Chen, A. Kolesnikov, S. Goel, T. Yun, P.-C. Chang, and A. Carroll, "Improving variant calling using population data and deep learning," *BMC Bioinformatics*, vol. 24, no. 1, May 2023, doi: 10.1186/s12859-023-05294-0.
[7] A. Khan and B. Lee, "DeepGene transformer: transformer for the gene expression-based classification of cancer subtypes," *Expert Systems with Applications*, vol. 226, Sep. 2023, doi: 10.1016/j.eswa.2023.120047.
[8] G. Yu, "Biomedical knowledge mining using GOSemSim and clusterProfiler," *YuLab-SMU*, 2022. Accessed: Jan. 03, 2024. [Online]. Available: https://yulab-smu.top/biomedical-knowledge-mining-book/enrichment-overview.html#gsea-algorithm
[9] T. Beck, T. Rowlands, T. Shorter, and A. J. Brookes, "GWAS central: an expanding resource for finding and visualising genotype and phenotype data from genome-wide association studies," *Nucleic Acids Research*, vol. 51, no. D1, pp. D986–D993, Jan. 2023, doi: 10.1093/nar/gkac1017.
[10] V. Patel and M. Shah, "Artificial intelligence and machine learning in drug discovery and development," *Intelligent Medicine*, vol. 2, no. 3, pp. 134–140, Aug. 2022, doi: 10.1016/j.imed.2021.10.001.
[11] Y. Zhang, S. Li, M. Xing, Q. Yuan, H. He, and S. Sun, "Universal approach to de novo drug design for target proteins using deep reinforcement learning," *ACS Omega*, vol. 8, no. 6, pp. 5464–5474, Feb. 2023, doi: 10.1021/acsomega.2c06653.
[12] S. Huang, S. Zheng, and R. Chen, "Multi-source transfer learning with graph neural network for excellent modelling the bioactivities of ligands targeting orphan G protein-coupled receptors," *Mathematical Biosciences and Engineering*, vol. 20, no. 2, pp. 2588–2608, 2022, doi: 10.3934/mbe.2023121.
[13] B. Johnston and P. D. Chazal, "Deriving ground truth labels for regression problems using annotator precision," *Applied Sciences*, vol. 13, no. 16, Aug. 2023, doi: 10.3390/app13169130.
[14] S. R. Khope and S. Elias, "Strategies of predictive schemes and clinical diagnosis for prognosis using MIMIC-III: a systematic review," *Healthcare*, vol. 11, no. 5, Feb. 2023, doi: 10.3390/healthcare11050710.
[15] M. Heath, T. H. Porter, and G. Silvera, "Hospital characteristics associated with HIPAA breaches," *International Journal of Healthcare Management*, vol. 15, no. 2, pp. 171–180, Apr. 2022, doi: 10.1080/20479700.2020.1870349.
[16] T. Edinburgh, S. J. Eglen, P. Thoral, P. Elbers, and A. Ercole, "Sepsis-3 criteria in AmsterdamUMCdb: open-source code implementation," *Gigabyte*, vol. 2022, pp. 1–7, Mar. 2022, doi: 10.46471/gigabyte.45.
[17] M. Sorkhi, M. R. Jahed-Motlagh, B. Minaei-Bidgoli, and M. Reza Daliri, "Learning temporal-frequency features of physionet EEG signals using deep convolutional neural network," *International Journal of Modern Physics C*, vol. 34, no. 4, Apr. 2023, doi: 10.1142/S012918312350047X.
[18] Z. Mustafa and H. Nsour, "Using computer vision techniques to automatically detect abnormalities in chest X-rays," *Diagnostics*, vol. 13, no. 18, Sep. 2023, doi: 10.3390/diagnostics13182979.
[19] F. Su *et al.*, "Deep learning-based synapse counting and synaptic ultrastructure analysis of electron microscopy images," *Journal of Neuroscience Methods*, vol. 384, Jan. 2023, doi: 10.1016/j.jneumeth.2022.109750.
[20] E. H. Houssein, R. E. Mohamed, and A. A. Ali, "Heart disease risk factors detection from electronic health records using advanced NLP and deep learning techniques," *Scientific Reports*, vol. 13, no. 1, May 2023, doi: 10.1038/s41598-023-34294-6.
[21] L. Alzubaidi *et al.*, "A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications," *Journal of Big Data*, vol. 10, no. 1, Apr. 2023, doi: 10.1186/s40537-023-00727-2.
[22] K. Rimal, K. B. Shah, and A. K. Jha, "Advanced multi-class deep learning convolution neural network approach for insect pest classification using TensorFlow," *International Journal of Environmental Science and Technology*, vol. 20, no. 4, pp. 4003–4016, Apr. 2023, doi: 10.1007/s13762-022-04277-7.
[23] A. Fernandez-Quilez, "Deep learning in radiology: ethics of data and on the value of algorithm transparency, interpretability and explainability," *AI and Ethics*, vol. 3, no. 1, pp. 257–265, Feb. 2023, doi: 10.1007/s43681-022-00161-9.
[24] Z. Ding, F. Sha, Y. Zhang, and Z. Yang, "Biology-informed recurrent neural network for pandemic prediction using multimodal data," *Biomimetics*, vol. 8, no. 2, Apr. 2023, doi: 10.3390/biomimetics8020158.
[25] V. H. Pereira-Ferrero, L. P. Valem, and D. C. G. Pedronette, "Feature augmentation based on manifold ranking and LSTM for

image classification," *Expert Systems with Applications*, vol. 213, Mar. 2023, doi: 10.1016/j.eswa.2022.118995.
[26] Y. Lu, D. Chen, E. Olaniyi, and Y. Huang, "Generative adversarial networks (GANs) for image augmentation in agriculture: A systematic review," *Computers and Electronics in Agriculture*, vol. 200, Sep. 2022, doi: 10.1016/j.compag.2022.107208.
[27] N. Zeng, H. Li, and Y. Peng, "A new deep belief network-based multi-task learning for diagnosis of Alzheimer's disease," *Neural Computing and Applications*, vol. 35, no. 16, pp. 11599–11610, Jun. 2023, doi: 10.1007/s00521-021-06149-6.
[28] M. Shi *et al.*, "Genetic-GNN: Evolutionary architecture search for graph neural networks," *Knowledge-Based Systems*, vol. 247, Jul. 2022, doi: 10.1016/j.knosys.2022.108752.
[29] S. Cao, J. Li, K. P. Nelson, and M. A. Kon, "Coupled VAE: improved accuracy and robustness of a variational autoencoder," *Entropy*, vol. 24, no. 3, Mar. 2022, doi: 10.3390/e24030423.
[30] A. Bulinski and D. Dimitrov, "Statistical estimation of the Kullback–Leibler divergence," *Mathematics*, vol. 9, no. 5, Mar. 2021, doi: 10.3390/math9050544.
[31] P. Ladosz, L. Weng, M. Kim, and H. Oh, "Exploration in deep reinforcement learning: a survey," *Information Fusion*, vol. 85, pp. 1–22, Sep. 2022, doi: 10.1016/j.inffus.2022.03.003.
[32] Y. Zhang, Z. Bai, and S. Imoto, "Investigation of the BERT model on nucleotide sequences with non-standard pre-training and evaluation of different k-mer embeddings," *Bioinformatics*, vol. 39, no. 10, Oct. 2023, doi: 10.1093/bioinformatics/btad617.
[33] N. S. Punn and S. Agarwal, "Modality specific U-Net variants for biomedical image segmentation: a survey," *Artificial Intelligence Review*, vol. 55, no. 7, pp. 5845–5889, Oct. 2022, doi: 10.1007/s10462-022-10152-1.
[34] R. Yang *et al.*, "Fabrication of liquid cell for in situ transmission electron microscopy of electrochemical processes," *Nature Protocols*, vol. 18, no. 2, pp. 555–578, Feb. 2023, doi: 10.1038/s41596-022-00762-y.
[35] S. Kang, H. Park, and J.-I. Park, "Identification of multiple image steganographic methods using hierarchical ResNets," *IEICE Transactions on Information and Systems*, vol. E104.D, no. 2, pp. 350–353, Feb. 2021, doi: 10.1587/transinf.2020EDL8116.
[36] H. Song, B. Zhao, J. Hu, H. Sun, and Z. Zhou, "Research on improved DenseNets pig cough sound recognition model based on SENets," *Electronics*, vol. 11, no. 21, Oct. 2022, doi: 10.3390/electronics11213562.
[37] P. ZHang, P. Wei, and Sh. Han, "CapsNets algorithm," *Journal of Physics: Conference Series*, vol. 1544, no. 1, May 2020, doi: 10.1088/1742-6596/1544/1/012030.
[38] M. Kang, S. Lee, D. Lee, and S. Kim, "Learning cell-type-specific gene regulation mechanisms by multi-attention based deep learning with regulatory latent space," *Frontiers in Genetics*, vol. 11, Sep. 2020, doi: 10.3389/fgene.2020.00869.
[39] B. Ni, Z. Liu, X. Cai, M. Nappi, and S. Wan, "Segmentation of ultrasound image sequences by combing a novel deep siamese network with a deformable contour model," *Neural Computing and Applications*, vol. 35, no. 20, pp. 14535–14549, Jul. 2023, doi: 10.1007/s00521-022-07054-2.
[40] U. A. Bhatti, H. Tang, G. Wu, S. Marjan, and A. Hussain, "Deep learning with graph convolutional networks: an overview and latest applications in computational intelligence," *International Journal of Intelligent Systems*, vol. 2023, pp. 1–28, Feb. 2023, doi: 10.1155/2023/8342104.
[41] O. Rubasinghe, X. Zhang, T. K. Chau, Y. H. Chow, T. Fernando, and H. H.-C. Iu, "A novel sequence to sequence data modelling based CNN-LSTM algorithm for three years ahead monthly peak load forecasting," *IEEE Transactions on Power Systems*, vol. 39, no. 1, pp. 1932–1947, Jan. 2024, doi: 10.1109/TPWRS.2023.3271325.
[42] C. Ren and H. Ren, "Prostate segmentation on magnetic resonance imaging," *IEEE Access*, vol. 11, pp. 145944–145953, 2023, doi: 10.1109/ACCESS.2023.3338746.
[43] S. Markidis, "Programming quantum neural networks on NISQ systems: an overview of technologies and methodologies," *Entropy*, vol. 25, no. 4, Apr. 2023, doi: 10.3390/e25040694.
[44] C. Comito and C. Pizzuti, "Artificial intelligence for forecasting and diagnosing COVID-19 pandemic: a focused review," *Artificial Intelligence in Medicine*, vol. 128, Jun. 2022, doi: 10.1016/j.artmed.2022.102286.
[45] H. Baek, W. J. Yun, S. Park, and J. Kim, "Stereoscopic scalable quantum convolutional neural networks," *Neural Networks*, vol. 165, pp. 860–867, Aug. 2023, doi: 10.1016/j.neunet.2023.06.027.
[46] Y. Li *et al.*, "Quantum recurrent neural networks for sequential learning," *Neural Networks*, vol. 166, pp. 148–161, Sep. 2023, doi: 10.1016/j.neunet.2023.07.003.
[47] L. Moro and E. Prati, "Anomaly detection speed-up by quantum restricted Boltzmann machines," *Communications Physics*, vol. 6, no. 1, Sep. 2023, doi: 10.1038/s42005-023-01390-y.
[48] N. Innan, M. A. Z. Khan, B. Panda, and M. Bennai, "Enhancing quantum support vector machines through variational kernel training," *Quantum Information Processing*, vol. 22, no. 10, Oct. 2023, doi: 10.1007/s11128-023-04138-3.
[49] M. B. Hastings, "Classical and quantum algorithms for tensor principal component analysis," *Quantum*, vol. 4, Feb. 2020, doi: 10.22331/q-2020-02-27-237.
[50] N. Friis, "Quantum decision making," *Quantum Views*, vol. 7, Feb. 2023, doi: 10.22331/qv-2023-02-07-72.
[51] J. Marino, "Predictive coding, variational autoencoders, and biological connections," *Neural Computation*, vol. 34, no. 1, pp. 1–44, Jan. 2022, doi: 10.1162/neco_a_01458.
[52] J. Earl, "The Belmont report and innovative practice," *Perspectives in Biology and Medicine*, vol. 63, no. 2, pp. 313–326, 2020, doi: 10.1353/pbm.2020.0021.
[53] V. Ashall, D. Morton, and E. Clutton, "A declaration of Helsinki for animals," *Veterinary Anaesthesia and Analgesia*, vol. 50, no. 4, pp. 309–314, Jul. 2023, doi: 10.1016/j.vaa.2023.03.005.
[54] P. Paraman and S. Anamalah, "Ethical artificial intelligence framework for a good AI society: principles, opportunities and perils," *AI and SOCIETY*, vol. 38, no. 2, pp. 595–611, Apr. 2023, doi: 10.1007/s00146-022-01458-3.
[55] F. Finotello, E. Calura, D. Risso, S. Hautaniemi, and C. Romualdi, "Editorial: multi-omic data integration in oncology," *Frontiers in Oncology*, vol. 10, Sep. 2020, doi: 10.3389/fonc.2020.01768.
[56] Z. Azhibekova *et al.*, "Using deep learning to diagnose retinal diseases through medical image analysis," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 14, no. 6, pp. 6455-6465, Dec. 2024, doi: 10.11591/ijece.v14i6.pp6455-6465.
[57] S. U. Maheswara Rao, K. Sreekala, P. S. Rao, N. Shirisha, G. Srinivas, and E. Sreedevi, "Plant disease classification using novel integration of deep learning CNN and graph convolutional networks," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 36, no. 3, Dec. 2024, doi: 10.11591/ijeecs.v36.i3.pp1721-1730.
[58] G. Alavani, J. Desai, S. Saha, and S. Sarkar, "Program analysis and machine learning–based approach to predict power consumption of CUDA kernel," *ACM Transactions on Modeling and Performance Evaluation of Computing Systems*, vol. 8, no. 4, pp. 1–24, Dec. 2023, doi: 10.1145/3603533.
[59] H. Bouraqqadi, A. Berrag, M. Mhaouach, A. Bouhoute, K. Fardousse, and I. Berrada, "PyFed: extending PySyft with N-IID federated learning benchmark," *Proceedings of the Canadian Conference on Artificial Intelligence*, Jun. 2021, doi: 10.21428/594757db.9c5550b5.

BIOGRAPHIES OF AUTHORS

Usman Tariq is an Associate Professor at Prince Sattam Bin Abdulaziz University in Saudi Arabia, earned his Ph.D. from Ajou University, South Korea, in 2010. With over $1 million in research funding, he has made significant contributions to wireless sensor networks, IoT systems, cybersecurity, and intelligent infrastructure, resulting in over 200 publications. A recognized expert, he has led international conferences and delivered over 50 keynotes and invited talks. He can be contacted at email: u.tariq@psau.edu.sa.

Irfan Ahmed recieved VCU NIRA Scholar, a Fellow of the American Academy of Forensic Sciences, and an Associate Professor of Computer Science at Virginia Commonwealth University (VCU). He is also a faculty fellow of VCU Cybersecurity Center. He can be contacted at email: iahmed3@vcu.edu.

Muhammad Attique Khan is currently an Assistant Professor with the Department of Computer Science, HITEC University, Taxila, Pakistan. He has above 200 publications that have more than 13000 citations and impact factor more than 600 with H-index of 68. His research interests include medical imaging, COVID19, MRI analysis, video surveillance, human gait recognition, and agriculture plants. He can be contacted at email: attique.khan@hitecuni.edu.pk.

Ali Kashif Bashir holds the position of Professor of computer networks and cybersecurity at the Manchester Metropolitan University in the United Kingdom. Throughout his career, he has presented more than 45 keynote speeches at international conferences, produced over 300 research articles, and successfully overseen multimillion-dollar projects worldwide. He can be contacted at email: a.bashir@mmu.ac.uk.