Review article

Prediction of diseases using transformers for gene expression dataset

Immaculate Mercy[1]* & M. Chidambaram[2]

[1] Department of Computer Science and Applications, Periyar Maniammai Institute of Science &Technology (Deemed to be University), Thanjavur, India
[2] Department of Computer Science, Rajah Serfoji Govt College, (Auto), Thanjavur, Affiliated to Bharathidasan University, Tiruchirappalli, India

*Corresponding author E-mail: immaculatemercy@pmu.edusss

Received: February 11, 2025
Corrected: June 10, 2025
Accepted: June 17, 2025

https://doi.org/10.15446/rcciquifa.v54n3.118814

SUMMARY

Introduction: Gene expression is a significant process which bridges the gap between information encoded within a gene and the final functional product of a gene. Therefore, evaluating gene expression is a vital process in terms of developing treatment and monitoring of disease, as diseases can result in unforeseen consequences. **Objectives**: Various studies have incorporated the Deep Learning (DL) concept with gene datasets for prediction of disease. However, prevailing DL methods are considered to be extremely ineffective in terms of accuracy and other aspects of disease prediction. Therefore, DL with transformers are applied for predicting the disease with gene dataset. As transformers are self-attention mechanisms, this permits the use of contextual information for any location in the input sequence and aids in capturing long range dependencies, which aids in delivering better accuracy and effective model for predicting disease. **Methods**: This paper focuses on the applications of gene expression with different DL algorithms along with transformers for prediction of diseases. Some of the applications of transformers with gene expression dataset include prediction of cancer such as lung cancer, stomach cancer and drug discovery, a focus of this paper. **Results**: Crucial analysis is undertaken by considering certain aspects such as architecture, dataset, process. This analysis helps in finding the aspects of the DL algorithms with and without transformers. Finally, gaps are identified through the analysis of transitional researchers and could be considered as future recommendations by overcoming the gaps that are intended to create promising work in this area. **Conclusion**: Despite having several advantages, there have been limited studies in terms of prediction of disease. Different applications such as cancer prediction, drug discovery are looked at in the present study. It has been observed that incorporation of transformers in DL algorithms were implemented more beginning from the year 2021; still there are a few complications which needs to be overcome for prediction of diseases. DL using transformers alongside gene datasets will be helpful in disease prediction. This study would aid researchers in their innovations in this area for enhancing the effectiveness and efficiency in prediction of disease.

Keywords: Gene expression; transformers; deep learning; gene datasets; applications.

RESUMEN

Predicción de enfermedades mediante transformadores para conjuntos de datos de expresión génica

Introducción: La expresión génica es un proceso importante que conecta la información codificada en un gen con su producto funcional final. Por lo tanto, evaluar la expresión génica es vital para el desarrollo de tratamientos y el seguimiento de enfermedades, ya que estas pueden tener consecuencias imprevistas. **Objetivos:** Diversos estudios han incorporado el concepto de aprendizaje profundo (ADP) con conjuntos de datos genéticos para la predicción de enfermedades. Sin embargo, los métodos de ADP predominantes se consideran extremadamente ineficaces en términos de precisión y otros aspectos de la predicción de enfermedades. Por lo tanto, se aplica el ADP con transformadores para la predicción de enfermedades con conjuntos de datos genéticos. Dado que los transformadores son mecanismos de autoatención, esto permite el uso de información contextual para cualquier ubicación en la secuencia de entrada y ayuda a capturar dependencias de largo alcance, lo que contribuye a obtener una mayor precisión y un modelo eficaz para la predicción de enfermedades. **Métodos:** Este artículo se centra en las aplicaciones de la expresión génica con diferentes algoritmos de ADP junto con transformadores para la predicción de enfermedades. Algunas de las aplicaciones de los transformadores con conjuntos de datos de expresión génica incluyen la predicción de cánceres como el de pulmón y el de estómago, y el descubrimiento de fármacos, temas centrales de este artículo. **Resultados:** Se realiza un análisis crucial considerando aspectos como la arquitectura, el conjunto de datos y el proceso. Este análisis ayuda a identificar los aspectos de los algoritmos de aprendizaje automático con y sin transformadores. Finalmente, se identifican lagunas mediante el análisis de investigadores en transición, las cuales podrían considerarse como recomendaciones futuras para superarlas y generar trabajos prometedores en esta área. **Conclusión:** A pesar de sus numerosas ventajas, los estudios sobre predicción de enfermedades son limitados. En el presente estudio se analizan diferentes aplicaciones, como la predicción del cáncer y el descubrimiento de fármacos. Se ha observado que la incorporación de transformadores en los algoritmos de aprendizaje automático se ha implementado con mayor frecuencia a partir de 2021; sin embargo, existen algunas complicaciones que deben superarse para la predicción de enfermedades. El aprendizaje automático que utiliza transformadores junto con conjuntos de datos genéticos será útil en la predicción de enfermedades. Este estudio ayudará a los investigadores en sus innovaciones en esta área para mejorar la eficacia y la eficiencia en la predicción de enfermedades.

Palabras clave: expresión génica; transformadores; aprendizaje profundo; conjuntos de datos genéticos; aplicaciones.

RESUMO

Predição de doenças usando transformadores para conjunto de dados de expressão gênica

Introdução: A expressão gênica é um processo significativo que preenche a lacuna entre a informação codificada em um gene e o produto funcional final de um gene. Portanto, avaliar a expressão gênica é um processo vital em termos de desenvolvimento de tratamento e monitoramento de doenças, visto que doenças podem resultar em consequências imprevistas. **Objetivos:** Diversos estudos incorporaram o conceito de Aprendizado Profundo (APL) com conjuntos de dados genéticos para a predição de doenças. No entanto, os métodos de APL predominantes são considerados extremamente ineficazes em termos de precisão e outros aspectos da predição de doenças. Portanto, APL com transformadores é aplicada para a predição de doenças com conjuntos de dados genéticos. Como os transformadores são mecanismos de autoatenção, isso permite o uso de informações contextuais para qualquer local na sequência de entrada e auxilia na captura de dependências de longo alcance, o que contribui para fornecer maior precisão e um modelo eficaz para a predição de doenças. **Métodos:** Este artigo foca nas aplicações da expressão gênica com diferentes algoritmos de APL, juntamente com transformadores, para a predição de doenças. Algumas das aplicações de transformadores com conjunto de dados de expressão gênica incluem a previsão de cânceres como câncer de pulmão, câncer de estômago e a descoberta de medicamentos, foco deste artigo. **Resultados:** Uma análise crucial é realizada considerando certos as-

pectos como arquitetura, conjunto de dados e processos. Essa análise auxilia na identificação dos aspectos dos algoritmos de DL com e sem transformadores. Por fim, lacunas são identificadas por meio da análise de pesquisadores em transição e podem ser consideradas recomendações futuras, superando-as, visando gerar trabalhos promissores nessa área. **Conclusão:** Apesar de apresentar diversas vantagens, estudos em termos de previsão de doenças são limitados. Diferentes aplicações, como previsão de câncer e descoberta de medicamentos, são analisadas no presente estudo. Observou-se que a incorporação de transformadores em algoritmos de DL foi implementada com mais frequência a partir de 2021; ainda existem algumas complicações que precisam ser superadas para a previsão de doenças. DL utilizando transformadores juntamente com conjuntos de dados de genes será útil na previsão de doenças. Este estudo auxiliará pesquisadores em suas inovações nessa área para aumentar a eficácia e a eficiência na previsão de doenças.

Palavras-chave: expressão gênica; transformadores; aprendizado profundo; conjuntos de dados genéticos; aplicações.

1. INTRODUCTION

Gene expression (GE) is a significant idea in molecular biology as it is frequently characterized by cell specificity and timing and regulation. The significance to control of gene expression for the progressive biologist is made noticeable by just seeing the nature of the field. Development refers to is utilized to define the coordination in space and time of various cellular undertakings, which includes migration, apoptosis, mitosis and differentiation [1].

Assessing the expression of gene is a fundamental process for developing treatments and monitoring of human diseases. Therefore, in order to increase the accessibility and approachability of the process, different DL methods can be used in this task [2]. The analysis of gene expression data is gaining attention due to its applications in prognosis, diagnosis of cancer and various other domains [3]. Gene functions, comprehending the regulation of gene, homology detection, subtypes of cells can be understood by employing gene expression data. Therefore, various clustering methods are employed for the analysis of gene expression data, however, the performance of the traditional techniques were considered to be poor and abridged as a result of high dimensionality of gene expression data. Hence, suggested study incorporates auto-encoder network, which aids in transforming high dimensions to low dimensions. Unsupervised feature selection methods were also employed in the suggested study for analyzing the gene expression data [4]. The researchers have employed various mathematical as well as statistical approaches in order to examine the gene expression data for various tenacities, including detection of informative pathways, enhanced classification of disease, and prediction of disease, drug discovery and personalized therapy. Therefore, different approaches have been developed with the aim to encounter these areas. Intricacy and high dimensionality of the GE data present significant challenges. Furthermore, the tools implemented for assessing the gene expression genowide was continuing to progress, which ultimately leads to enhanced accuracy in values of GE. Gene expression patterns are detected using RNA-seq instead of employing DNA-micro arrays [5]. Consequently, technology development required new mathematical and statistical approach with the aim to examine the heterogeneous data. Further difficulty includes, the interacting factors such as dietroit factors, asbestos can interacted and influence the genes associated with the specific cancer [6]. While various ML techniques exist for application of gene expression, DL is more effective and aids in delivering

precise prediction based on gene expression data. In DL, different ways of networks are employed for prediction however, transformers are utilized in minimum studies. Therefore, the present study focuses on projecting DL and DL with transformers for predicting the disease with gene datasets.

Transformer entirely relies on self-attention to handle long-range dependencies of sequence data. It was first proposed for NLP tasks [7, 8]. In order to apply transformers on image data, Dosovitskiy *et al.* [9] split an image into patches and treated them as tokens. Hence, a pure transformer can be adopted. Further it has been identified that, the transformer has high computational efficiency and scalability to train models of large sizes with more than 100 B parameters [10], which makes it even more compatible and vital for various fields. The generative model designed to improve disease outcome prediction from longitudinal electronic health records (EHRs). The objective is to utilize a transformer architecture that pertains on extensive HER data to predict future disease outcomes based on patients data, enhancing the model's ability to capture complex temporal relationships. The data enhancing employs an encoder-decoder framework, where the encoder processes input embedding and the decoder generates predictions for future clinical visits, effectively leveraging the self-attention mechanism to capture long range dependencies in the data [11].

Transformers are described as network architectures which are depended on multi head self-attention mechanism (SAM) that permits capturing the lengthy-range dependencies between items a sequence [12]. Transformer data are used for processing the sequential data, which includes genomic sequence, NL data, time series data or acoustic signals. Transformer Neural network (TNN) is considered as appealing choice of network for analysis of gene expression, owing to its capability to conjointly appear the information from several demonstration subspaces at various places in genomic data. Gene transformer is [13] uses multi-head self-attention modules with the aim to encounter the difficulty of the high dimensional GE by identifying the appropriate biomarkers across multiple subtypes of cancer. Similarly, suggested study has employed multi-omic transformer so as to distinguish complex phenotype depending on the 4 omic data types [14]. Correspondingly, recommended study implemented transformer based fusion network (T-FN) [15] or analysis of cancer survival by incorporating genomic data and pathological images. TFN module permitted the researchers to leverage the intramodality relations between the patches in various domains for multiple pathological slides. Similarly by leveraging a multi head self-attention mechanism, the model effectively captures complex gene interaction and long range dependencies, enhancing the identifications of relevant biomarkers. The outcome demonstrates the DeepGene outperforms traditional classification algorithm across various metrics, achieving higher accuracy and fewer misclassification errors. This approach underscores the potential of transformer architectures in advancing precision [16].

It has been revealed that, transformer models have demonstrated increased robustness and sturdiness more than just CNN and RNN models which delivered competitive performance on standards with different formats of data. One of the advantages of employing Transformers is that, the self-attention mechanism permits to use contextual information for any position in the input sequence and aids in capturing long range dependencies when compared to other models and further, it allows higher parallelization when compared to RNN model [17]. However, there are few drawbacks of Transformers which needs to be considered, that includes, need of huge amount of data, therefore the performance of the model can be sub-

standard when compared to other NN models for genetic data with minimum number of input samples [18].

Therefore the objectives of the paper include reviewing the research works related to predictive DL classifiers for gene expression dataset, analyzing the existing DL algorithm using transformers for the prediction of diseases, discussing the application of gene expression in different aspects, addressing challenges in current methodologies, and identifying research gap. Additionally, this review aims to bridge the gap between emerging transformer based model and their practical application in bioinformatics, summarizing the current state of research while providing actionable insights for future studies, such as integrating transformers with data and high throughput techniques in genomics.

1.1. Paper organization

The paper is organized in the following way. Section 2, describes the survey methodology of the paper, overall vision about gene expression is focused on section 3. Evolution and applications of DL classifiers for gene expressions is elucidated in section 4, different DL based algorithms for gene expression datasets is explained in section 5, comparative analysis is depicted in section 7, critical analysis is illustrated in section 7, research gaps are depicted in section 8, future recommendations in section 9 and conclusion is depicted in section 10.

2. SURVEY METHODOLOGY

The survey methodology is divided into 5 stages/steps for reviewing the paper which includes review plan, research questions identification, data sources, search criteria and finally inclusion/exclusion criteria. In-depth description of the above mentioned steps are described as follows.

Initially, the proposed survey was done systematically by reviewing the papers such as identification of possible research questions, identification of data sources, effective search strings/criteria, and inclusion and exclusion criteria for articles. To perform this study there is a need for some pre-planning. The identification of relevant publications in DL with transformers has been considered for this survey that proved helpful for the proposed survey.

After the process of preplanning, research questions were framed accordingly. Table-1 shows the possible research questions the authors have identified in carrying out the proposed survey along with their objectives.

Table 1. Research Questions and Objectives

Q.No	Research Questions	Objectives
RQ 1	Why to choose transformers for predicting disease over other models	To get an overview of Transformers and its advantages for predicting the disease.
RQ 2	What are the solutions that were beneficial in the past	Through this question, more related literature could be identified related to that area.
RQ 3	What do past survey paper have concluded?	It aimed in comparing the proposed survey with existing papers

After framing research questions with objectives, appropriate sources should for the paper were selected. Therefore, Broad number of reputed and trusted literature have been considered for writing a detailed survey on DL using transformers. Standard digital libraries and databases like IEEEXplore, Springer, Frontier, Science Direct (Elsevier), Bioinformatics, IET,

Wiley online library, Research gate etc. Electronic data sources are also recommended for the literature survey such as articles.

Further, the search criteria was carried out by considering specific keywords like''Gene Expression", "Transformers", ''Deep Learning using transformers'', "Gene Expression Dataset", "Applications of gene expression using DL with transformers", "Hybrid transformers", and other related keywords. Majorly considered the papers that provide a survey on DL model as well as the centralized model in Transformers.

Eventually, after fetching many papers, the papers that were relevant to the topic (by analyzing title, abstract, introduction, and conclusion) were only considered. The papers that included an introduction to DL and a centralized system and which have the same topic were taken into consideration. To make the survey interactive and effective, we preferred to include recent papers of the year 2019, 2020, 2021, 2022, 2023 and early access articles. The filtering of papers and articles are based on the parameters mentioned above. Finally, we identified and focused on some papers having good citations. The procedures followed in studying these approaches has been shown in the Figure 1.

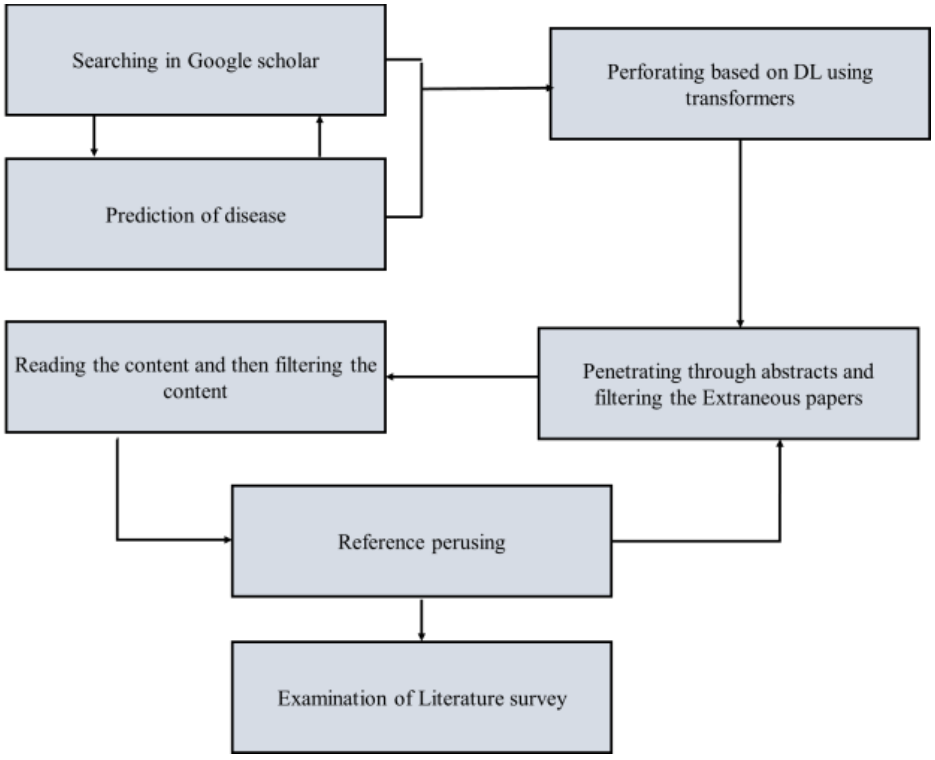

Figure 1. Survey methodology

3. OVERALL VISION ABOUT GENE EXPRESSION

Gene expression (GE) is the practice by which the information comprised in DNA is transmuted into instructions for creating the proteins or additional molecules [19]. This process encompasses, transcription of DNA into mRAN (messenger RNA), then subsequently translated into proteins. The GE Analysis (GEA) is utilized to evaluate the directive of genetic modifications that occurs under specific circumstances, in tissue or a single cells [20]. A component of GE quantification is a contrast of the reads that are subsequent, associated with the amount of base pairs sequenced from a fragment of DNA to a familiar transcriptome or genomic

source. Therefore, predominant techniques for assessing GE include NGS methods (Next Generation Sequencing) and DNA microarray. The DNA microarray technique uses a 2D array with infinitesimal spots, whereas the NGS technique delivers high speed, scalability and throughput analysis [21, 22]. RNA-seq is considered as NGS technique which involves the transformation of molecules of RNA to cDNA (complementary DNA). However, when compared to DNA microarrays, RAN-seq delivered various advantages which includes greater resolution, higher dynamic range, increased sensitivity and greater specificity. Transcriptome for any species can be examined by employing RNA-seq with the aim to control the sum of RNA at a precise time [23, 24].

GEA needs different computational approaches in order to comprehend the regulation of genes and the part in functioning of tissues and cells. Therefore, various ML and DL methods have been employed to attain the perceptions on how disparities in regulatory regions and genes results in phenotypic variations namely, health, wellness and traits [25]. Further, early computational methods for GEA has been depended on different ML and DL approaches for predicting the function and structure of genomic components [26]. Moreover, feature engineering is considered as one of the methods, which is used for handling the difficulty of high dimensionality and moderately smaller number of samples in GE data.

3.1. Gene Expression Data

GEA is the practice of detecting the amount of transcripts exist in specific tissue or cells in order to assess the level of expressed genes [27]. Some of the gene GED are demonstrated.

3.1.1. Microarray data

Microarray data are attained via a lab procedure, in which a sequence of DNA is comprised in a tool which comprises of 2D array with hundreds and thousands of microscopic spots [28]. The Microarray data are employed with the aim to find the DNA or RNA [29]. Microarray data permits for understanding of cellular process for genome-wide expression profile, which are based on particular conditions or diseases such as cancer.one of the advantages of employing DNA microarray is that, they permit the evaluate the expressional level of 1000 genes, however, there are few drawbacks, which includes poor accuracy and specificity.

Training transformer model for gene expression analysis present presents several analysis present various implementation challenges, particularly in computational requirements. These model are memory-intensive and requires significant GPU resources to manage high-dimensional datasets effectively. Strategies such as gradient check pointing can help optimize memory usage. Effective data pre-processing is also critical for improving model performance, with key steps including normalization techniques like quintile normalization and log transformation to stabilize variance, as well as feature selection methods such as variance thresholding and statistical tests to identify relevant genes. Handling missing data through imputation and applying dimensionality reduction techniques like PCA can further enhance dataset robustness. Addressing these challenges and pre-processing steps will provide valuable insights into the practical application of transformer models in bioinformatics. This version is more concise while preserving the original meaning and key details.

3.1.2. RNA-Seq Data

RNA-Seq method fit into NGS techniques as they possess the ability for rapid profiling [30]. RNA-Seq is utilized for measuring the expressions of gene ad aids in scrutinizing the variations in GE. The applications of RNA-Seq is according to the capacity and examine all molecules of RNA in tissues and cells, which includes mRNA, miRNA, siRNA or rRNA [30]. One

of the significant qualities of RNA seq- includes huge dynamic range and high resolution, which resulted in huge volume of attained data and contributed to significant improvements in transcriptomics research [29].

4. SIGNIFICANT EVOLUTION AND APPLICATIONS OF DEEP LEARNING (DL) CLASSIFIERS FOR GENE EXPRESSION

Different methods have been used in the past decades for gene expression, which includes both ML and DL methods, however, in most studies, DL techniques are employed as DL possess unbelievable potential for backing medical and paramedical specialists by minimizing the human error rate, by aiding in diagnosing cancer and assist in analyzing the difficult data. Further, different applications are employed for gene expressions which includes classification of cancer. Therefore, suggested paper has employed DL methods for classification of cancer by employing CNN, as CNN is appropriate for analyzing the variety of unstructured data, in addition GE data consist of large number of genes and numerous researchers analyzed and evaluated the classification of cancer by utilizing different DM, ML and statistical approaches [31, 32]. Though, ML methods have attained decent accuracy in terms of classification, there are some of the issue like inability of the ML algorithms to employ unstructured data has restricted the utility of ML algorithms for classification tasks. Therefore, DL algorithms are used for classification as they possess the ability to comprehend complex data. Hyper parameters are essential setting that govern the learning process and significantly influence model performance, with learning rate, this parameters controls how much to change the model in response to the estimated error each tm eth model weights are updated. The capability of the DL algorithms aids in identifying the cancer. Further, hybrid DL model is employed in the suggested study and the hybrid model includes LS-CNN (Laplacian score-CNN), which aided in classification of cancer data [33]. The paper investigates the impact of activation function on transformer based models by utilizing rational activation functions (RAFs), which can adaptively learn optimal activation functions from data. The influences model performance and can reduce the topological complexity of input data. Their experimental demonstration that the RAF-based Transformer model (RAFT) outperforms traditional fixe activation functional models, achieving better result on GLUE and SQuAD [34]. Well the epochs of 0.001 and 2000 by achieving 99.4% accuracy, refers to the number of complete passes through the training dataset. Selecting an appropriate number of epochs is crucial since too few may lead to under fitting, while too many can result in over fitting.

Similarly, search results indicates that DL based methods have generally outperformed conventional ML methods, with many approaches employing MLP or CNN networks achieving test accuracies upwards of 90%. However the performance of current methods is sensitive to various parameters, necessitating future improvements for generalization and robustness. Other limitations includes a lack of interpretability and limited integration with other data types. Table-2 shows different NN which can be used.

Table 2. Different Neural networks

Network	Features	Pros	Cons
CNN	No manual extraction was needed as CNN learn features themselves image image/data and perform the extraction directly from images [35]	The model learns rapidly and steadily Performance of the model is good Possess strong ability to extract the features	Label is needed for classification The time required for training is long
RNN	Possess the ability to learn sequences and the weights are shared across all steps and neurons [36]	It learns sequential events It performs significantly better performance	Issues to vanishing gradient is possible
AE	It is significantly employed reduction of dimensionality and extraction of features, the number of the inputs will be equal to number of output.	Labeled data is not required.	Training may be disappeared It require a pre-training step
DBN	This is considered to be a unidirectional connection and it is performed for both supervised and unsupervised [37]	Over-fitting can be minimized It is effective in terms of usage of hidden layers [38]	Process can be computationally expensive due to initialization process [39]
TNN (Transformers Neural Network)	Transformers are NN which learns context and understanding via sequential data analysis [40]	Possess the capability to pass various words via NN concurrently.	Sometimes, Transformers can be difficult to interpret and comprehend, since they are trained end to end [41]

Similarly, recommended study focused on classification of cancer using gene expression data. Different datasets employed in the suggested paper includes prostate tumor, colon tumor, DLBCL and leukemia. Classification of gene expression data divides the cancer samples from health samples. However, the ML techniques were not capable enough to perform cancer classification due to the small number of samples with huge number of samples in GE data. Hence DL algorithm has been employed by the existing paper for classification of cancer. Different DL algorithms such as DNN, RNN, CNN and I-DNN were employed, in which I-DNN delivered better outcome than the rest of the algorithms [42]. The choice optimization algorithm affects how quickly and effectively a model learns. Different optimizers have unique strengths and weakness depending on the specific characteristics of the datasets achieved 90% in test accuracy. Correspondingly, a scalable and robust DL method is used in the existing study. The model is defined as GNE model which has been aided to perform a gene network embedding. Further, gene interactions can be predicted precisely by incorporating ML classifiers [43]. Likewise, a regression based MLP- SAE model was incorporated for predicting the gene expression from SNP genotypes [44]. A semi-supervised DL method was employed for target gene expression inference. Therefore, GAN model has been employed. Existing model incorporated the profiles with only landmark genes into the training process. The datasets incorporated in the recommended paper were GEO dataset, GTEx dataset [45]. Though existing methods delivered better outcome for prediction of diseases, it still lagged in delivering effective outcome for prediction and classification of diseases.

4.1. Transformers

The transformer model represents a watershed moment in the evolution of deep learning models. Distinct from conventional sequence transduction models, which typically involve recurrent or convolutional layers, the transformer model solely harnesses attention mechanisms, setting a new precedent in tasks such as machine translation and natural language processing (NLP). The principal component of a transformer model is the attention mechanism, and it comes in two forms: self-attention (also referred to as intra-attention) and multi-head attention. The attention mechanism's core function is to model interactions between different elements in a sequence, thereby capturing the dependencies among them without regard to their positions in the sequence. In essence, it determines the extent to which to pay attention to various parts of the input when producing a particular output.

Self-attention mechanisms operate by creating a representation of each element in a sequence that captures the impact of all other elements in the sequence. This is achieved by computing a score for each pair of elements, applying a Softmax function to obtain weights, and then using these weights to form a weighted sum of the original element representations. Consequently, it allows each element in the sequence to interact with all other elements, providing a more holistic picture of the entire sequence. The multi-head attention mechanism, on the other hand, is essentially multiple self-attention mechanisms, or heads, operating in parallel. Each head independently computes a different learned linear transformation of the input, and their outputs are concatenated and linearly transformed to result in the final output. This enables the model to capture various types of relationships and dependencies in the data. The self-attention mechanism in transformers significantly improves performance in handling gene expression data by enabling the model to capture long range dependencies and relationships between genes. Unlike, the traditional methods that process sequences in a sequential manner, self-attention allows for parallel processing, which enhances computational efficiency and scalability. This capability ensures that relevant interactions among gene are effectively modelled, leading to more accurate predictions and insights in gene expression, by advancing the field of genomics and personalized medicine by the self-mechanism. The capability ensures that relevant interactions among genes are effectively modelled, leading to more accurate prediction and insights in gene expression. Various existing studies have focused on employing transformers for attaining results. The training process for transformer model can be source intensive, often requiring substantial GPU memory and time. For example, the model with millions of parameters can demand upward of 40 GB of GPU memory due to their complex architecture and high dimensional data processing requirements. As sequence length increases, memory usage grown quadratically, necessitating careful management of batch size and input dimensions to optimize GPU utilizations. Hence, various existing studies have focused on employing transformers for attaining effective results.

Hence, the suggested paper has used a transformer-based neural network framework for prokaryotic genome annotation, primarily focusing on Escherichia coli. The study emphasized that a substantial part of the model's subunits or attention heads were attuned to identify transcription factors and characterize their binding sites and consensus sequences. With the specialization of the attention heads occurring automatically, it has been believed that transformer models to be of high interest towards the creation of explainable neural networks in this field. Further, this method opened the door to understanding well-known and possibly novel elements involved in transcription initiation [46].

Cell type identification by analyzing scRNA-seq data is mostly limited by time-consuming and irreproducible manual annotation. As scRNA-seq technology scales to thousands of cells per experiment, the exponential increase in the number of cell samples makes manual annotation more difficult. Therefore, suggested paper focused on Cell type identification by analyzing scRNA-seq data is mostly limited by time-consuming and irreproducible manual annotation. As scRNA-seq technology scales to thousands of cells per experiment, the exponential increase in the number of cell samples makes manual annotation more difficult. Hence, a scTransSort, a cell-type annotation method pre-trained with single-cell transcriptomics data has been used. The scTransSort incorporates a method of representing genes as gene expression embedding blocks to reduce the sparsity of data used for cell type identification and reduce the computational complexity. The feature of scTransSort is that its implementation of intelligent information extraction for unordered data, automatically extracting valid features of cell types without the need for manually labelled features and additional references. Experiments were conducted using a large number of scRNA-seq datasets from different species, different tissues, and different platforms. All scRNA sequence datasets used in this paper are published and publicly available datasets from multiple high-quality reports and the Gene Expression Omnibus (GEO) [47].

Similarly, a Transformer-based deep learning method, GPTransformer, that uses genotypic and phenotypic data to predict FHB (Fusarium head blight) severity and DON (Deoxynivalenol) levels in a two-row barley population, in addition, effect of feature selection on genomic prediction using mutual information was investigated and examine the biological relevance of the top markers identified by the mutual information method. Further, the Transformer network is trained using a graphical processing unit (GPU). As the internal mechanism of the Transformer creates a four dimensional matrix of size (batch size, number of heads, number of markers, number of markers) at certain point, which requires a large amount of GPU memory, the feature selection process also helps us to solve the GPU memory issue of the Transformer. This study suggests the potential of the Transformer based method as an alternative to the popular BLUP model for genomic prediction of complex traits such as FHB or DON, having performed equally or better than existing machine learning and statistical methods [48].

The suggested study has incorporated Hierarchical Multi Head Self Attention (H-MHSA) to make self-attention computation in transformer flexible and efficient. Specifically, we first split an image into patches, each of which is treated in the same way as a token. Instead of computing attention across all patches, we further group patches into small grids and compute attention within each grid. This step captures local relationships and yields more discriminative local representations. Then, merge these small grids into larger ones and compute attention within each new grid by viewing small grids at the preceding step as tokens. In this way, we essentially capture feature relationships in the larger region. This process is iterated to reduce the number of tokens gradually. Throughout this procedure, our H-MHSA computes self-attention in the increasing region sizes step by step and naturally models the global relationship in a hierarchical manner. Since each grid at each step only has a small number of tokens, we can reduce the computational/space complexity of vision transformer dramatically. We empirically observe that this strategy brings us better generalization results [49].

Though, there are various DL algorithms which is used for gene expression, DL with transformer provides additional advantages which makes the prediction accurately and the model effectively. Some of the reasons of implementing TNN over other approaches includes,

they can be more precise as they can comprehend the relationship between the sequential elements which are far from each other and they are quick at handling a sequence since lot of attention is paid for its most vital parts.

The study found that while Vision Transformers (ViTs) achieved a high accuracy of 95.3% for lung cancer classification, CNN like ResNet50 performed comparably at 94.5%, suggesting that both architectures can be effective but may excel under different conditions, particularly with varying dataset sizes. However, the inherent complexity of transformers complicates performance assessment; their ability to capture long range dependencies offers advantages in understanding intricate patterns, yet this benefit is not clearly delineated against the strength of CNN and RNNs in handling local features. Thus, while transformers shows the efficiency, the transition from the traditional models raises questions about the clinical significance of the performance improvements [50].

Therefore, transformers are considered as a type of ANN which is utilized to unravel the concerns of transformation or transduction of input sequences into output sequences in DL applications. One of the significant merits of employing the transformer architecture over other NN is that, transformers are efficient and unlike RNN, the input data do not have the need to be processed sequentially which permit the parallelization [51]. The attention mechanism is defined mathematically as:

4.1.1. Scaled Dot Product Attention:

$$Attention\ (Q, K, V) = softmax(\frac{QK^t}{\sqrt{d_o}})V$$

Where Q the matrix of queries, K is is the matrix of Keys, V is the matrix of values, and d_k is the dimension of the key. The attention mechanism computes a score for each query against all keys, normalized these scores using Softmax, and then uses these probabilities to weigh the corresponding values. This allows the model to focus on relevant parts of the input sequence when generating outputs.

4.1.2. Multi-Head Attention:

The multi-head attention function allows the model to jointly attend to information from different representation subspaces at different positions:

$$MultiHead(Q, K, V) = Concat(head_1, \ldots., head_h)W^O$$

Where each head is computed as:

$$head_i = Attewntion(QW_i^Q, KW_i^k, VW_i^v)$$

Multi head attention allows the model to attend to different parts of the input sequence simultaneously. By having multiple heads, the model can capture various relationships and dependencies within the data more effectively than a single attention mechanism.

4.1.3. Feed forward Neural Network:

Each position is processed independently through a feed forward neural network:

$$FFN(x) = ReLU(xW_1 + b_1)W_2 + b_2$$

The feedback network processes each position independently after the attention mechanism has been applied.

4.1.4. Layer Normalization:

Layer normalization is applied as follows:

$$LayerNorm(z;\gamma,\beta) = \gamma\frac{(z-\mu)}{\sigma} + \beta$$

Where μ and σ are the mean and standard deviation of the layer inputs. The layer normalization helps stabilized training by normalizing inputs across features rather than across batches. This ensures consistent performance across different layers and helps mitigate issues related to internal covariate shift.

4.1.5. Positional Encoding:

The position encoding used to inject information about the position of tokens in the sequence is defined as:

$$pk, 2i = \sin\left(\frac{k}{10000^{2i/d}}\right), \qquad pk, 2i+1 = \cos\left(\frac{k}{10000^{2i/d}}\right)$$

Since, transformers do not have a built in notion of sequence order, positional encoding injects information about token position into their embedding. This helps the model understanding the relative positions of tokens in sequences, which is crucial for task like language modelling or sequence prediction. The transformer outperforms traditional model by providing a global contextual understanding, enhancing predictive power, ad offering improved interpretability for disease prediction, so transformer model is efficient than traditional disease prediction model.

Self-attention mechanism is considered as the core of transformer, which learnt the dependencies of global features. Limitations of the self-attention addresses the existing DL techniques with the aim to model the genomics of the data. Unlike CNN, self-attention possess the capability to unordered the input which includes genes in gene expression data [23]. Table 3 depicts the utilization of transformers in existing studies.

Table 3. Utilization of Transformers in Existing Studies

Sl. No	Purpose	Findings	Reference
1.	A transformer-based neural network framework for prokaryotic genome annotation, primarily focusing on *Escherichia coli*	The study emphasized that a substantial part of the model's subunits or attention heads were attuned to identify transcription factors and characterize their binding sites and consensus sequences. With the specialization of the attention heads occurring automatically, it has been believed that transformer models to be of high interest towards the creation of explainable neural networks in this field	[46]
2.	To incorporate scTransSort method for representing genes as gene expression embedding blocks to reduce the sparsity of data used for cell type identification and reduce the computational complexity	A scTransSort, a cell-type annotation method pretrained with single-cell transcriptomics data has been used. The scTransSort incorporates a method of representing genes as gene expression embedding blocks to reduce the sparsity of data used for cell type identification and reduce the computational complexity. The feature of scTransSort is that its implementation of intelligent information extraction for unordered data, automatically	[47]

		extracting valid features of cell types without the need for manually labeled features and additional references.	
3.	To predict FHB severity and DON levels in a two-row barley population, in addition, effect of feature selection on genomic prediction using GPTransformer	The effect of feature selection on genomic prediction using mutual information was investigated and examine the biological relevance of the top markers identified by the mutual information method. Further, the Transformer network is trained using a graphical processing unit (GPU). As the internal mechanism of the Transformer creates a four dimensional matrix of size (batch size, number of heads, number of markers, number of markers) at certain point, which requires a large amount of GPU memory, the feature selection process also helps us to solve the GPU memory issue of the Transformer	[48]
4.	In order to make self-attention computation in transformer flexible and efficient by employing Hierarchical Multi Head Self Attention (H-MHSA)	Specifically, we first split an image into patches, each of which is treated in the same way as a token. Instead of computing attention across all patches, we further group patches into small grids and compute attention within each grid. This step captures local relationships and yields more discriminative local representations. Then, merge these small grids into larger ones and compute attention within each new grid by viewing small grids at the preceding step as tokens. In this way, features can be captured in larger region. Further, this process is iterated to reduce the number of tokens gradually. Throughout this procedure, the H-MHSA computes self-attention in the increasing region sizes step by step and naturally models the global relationship in a hierarchical manner. Since each grid at each step only has a small number of tokens, we can reduce the computational/space complexity of vision transformer dramatically.	[49]

Usually, a transformer is a NN architecture which exploits the idea of attention as well self-attention in a stack of decoders and encoders. Author, in the suggested paper has employed an architecture which has 6 encoders and 6 decoders. Initially, the input vectors are taken by the encoder and practice them with self-attention layer afore passing it to the forward NN. In standard architecture, each encoder will be comprised of 2 segments which includes Self-attention layer (SAL) and feed forward NN. Correspondingly, the every encoder is comprised of 3 segments which includes SAL Decoder Attention layer (DAL) and feed forward NN. Once the network of the first encoder has been done, then the outcome is passed to the next encoder and son on, until the final encoder sends the information to the decoders and a similar process supervenes [51].

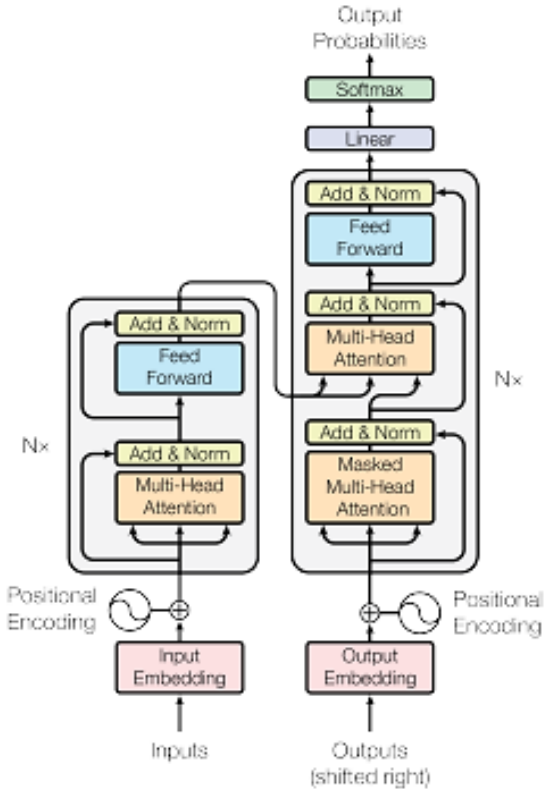

Figure 2. Transformer architecture [51].

Figure 2 shows the architecture of transformer employed in the recommended paper. Transformers 19, 23 and 24 received lot of consideration for biological sequence data pre-processing. These transformed are considered as a prevailing learners of data which are sequential. Moderately attributable to the employment of the SAM, which take cares the pairwise interdependencies among the elements, which in turn, are sequence in nature. Lately, the transformers are also assessed as the prevalent selections for self-supervised learning on biological sequences [52]. Transformers are DN (deep networks) with consist of self-attention mechanism in each layer, which permits attaining numerous enhancements pertaining to RNN and CNN models [53]. Like transformer, suggested paper has employed R-Transformer, which utilizes the advantages of both RNNs and the multi-head attention mechanism. This model aids in capture both local structures and global long term dependencies in sequences without any use of position embedding [54].

5. DIFFERENT DL BASED ALGORITHMS FOR GENE EXPRESSION DATASETS

5.1. Prediction of Disease

Author, have identified the genes demonstrated cancer tissues or 32 normal tissues, which was employed to examine the functional differences between the genes widely and complete outcome of the gene expression was exemplified rarely. The recommended outcome was aided in detecting the landscape of gene expression and comprehending the influence of gene expression and the microenvironment of cancer [55].

Identifying disease-related miRNA is essential for disease diagnosis and treatment. However, traditional biological experiments are highly uncertain and time-consuming. Hence, advanced intelligent computational models are needed to address this problem. Therefore, a dual-channel transformer graph model, named DCTGM, to learn multi-scale representations for miRNA-disease association prediction. Specifically, DCTGM includes a transformer encoder (TE) and GraphSAGE encoder (GE). The TE intensely captures the important interaction information between miRNA-disease pairs, and the GE aggregates multi-hop neighbor information of miRNA-disease association heterography to enrich node features. Then, an attention module is proposed to aggregate the dual-channel interactive representations, and we adopt a multi-layer perceptron (MLP) to predict the miRNA-disease association scores. For gene expression dataset, statistical analysis is crucial in identifying differentially expressed genes. Preprocessing steps, typically involves normalization to remove systematic biases while preserving biologically relevant variations. Techniques such as log transformation and quantile normalization are often employed to ensure comparability across samples. Additionally, methods like the Two-stage Poisson model can be utilized to account for over-dispersion in RNA-seq data, enhancing the accuracy of differential expression analyses [56].

Classification of cancer using gene expression is considered as one of the tremendously stimulating aspects, as the dimensionality and complexity of the data is high. Existing classification models were typically depend on samples which was gathered from a distinct tissue type and aided in performing a pre-requisite of gene feature selection, in order to prevent the processing the full set of genes. However, these technologies lacked in captivating benefits of genome-wise subsequent generation sequencing technologies, which delivered an image of the entire transcriptomics. Therefore, the recommended model aided DL structure for diagnosis of cancer by evolving a multi-class cancer classifier which was according to entire transcriptome GE that was gathered from numerous types of tumor which covered the multiple organ sites. Hence, suggested study employed a CNN architecture, termed as Gene Xpression Network (GeneXNet). This model was specifically designed for addressing the complicated nature of gene expressions. The recommended GeneXNet model provided abilities of identifying the alterations in genetics, driving the progression of cancer by learning genomic similarities across various kinds of tissues without craving any precondition of gene feature selection. Due to these reasons, GeneXNet model was capable of classifying the multi-tissues of cancer. From, the experimental outcome it was identified that recommended model delivered accuracy rate of 98.9% on human samples which represented 33 various type of cancer tumor across 26 organ sites [57].

A neuro-fuzzy approach was employed for examining the expression of gene data from microarray experiments. Analyzing the expression of gene led to detection and classification of cancer, which facilitated appropriate treatment selection and improvement of drug. The existing technique was tested on 3 benchmark cancer gene expression datasets. From the experimental outcome, it was identified that, neuro-fuzzy technique could be employed as an effective tool for computation of microarray data examination. The neuro classification system was according to the built-clustering algorithm, which got acknowledgement rate than the other classifiers. GE dataset for liver cancer was used in the existing study. The quality of the images were trained and tested for each dataset in order to enhance the quality of the genes [58].

Molecular subtyping of cancer is considered as a crucial stage towards more individualized therapy and delivers more significant biological understandings into cancer assortment.

Though in last decade, gene expression signature was widely demonstrated as one of the effective approaches for classification, pervasive implementation has been inadequate by batch effects, metamorphoses in platform and the trouble to classify the individual patients' samples. Therefore, a supervised method for classification of cancer was used. The model incorporated was DeepCC model (deep cancer subtype classification). DeepCC was based on DL. DeepCC classifiers were used for achieving higher specificity, sensitivity, accuracy when compared to the existing models like RF, SVM, XGBoost, GBM and multinomial LR algorithms. Robustness of the DeepCC model was demonstrated on simulation analysis. Moreover, DF (Deep Features) learnt by DeepCC apprehended the biological features, which was based on biological characteristics that was related to different molecular subtypes, which enabled more compressed distribution of subtypes. Hence, recommended paper aided in reducing the amount of unclassifiable samples earlier. Therefore, it was identified that, DeepCC model was beneficial for classification, which was platform independent, robust of omitted data and can be employed for single sample calculation which enabled clinical application of molecular subtyping of cancer [59].

5.1.1. Lung cancer

A different method for attaining tailored biomarker scores explained the prominence of every gene in detecting the subtype of the cancer for every sample by utilizing an AM. This modified significance score was employed for PCA (post classification analysis). NMF and PCA were both performed with the aim to identify the developing patterns within the scores of biomarkers. Nevertheless, the NMF and PCA were independent of task which are associated to classification, as there was no necessity for an end to end technique which could be implemented openly for countless real time applications. Further PET images contained information which were metabolic as they have been utilized for sub-classification of lung cancer. Nevertheless, it was challenging to mine massive PET images of carcinoma patients for DL associated applications [60].

Likewise, different type of cancer was classified according to the RNA sequence GE data by implementing DL technique which was based on BPSO-DT (Binary Particle with Decision Tree) and CNN. Different types of cancers, which could be identified were BRCA (Breast Invasive Carcinoma), LUAD (lung adenocarcinoma), UCEC (Uterine Corpus Endometrial Carcinoma), and LUSC (Lung Squamous Cell Carcinoma). The recommended model comprised of 3 stages. The first phase includes pre-processing. This step aided in optimizing the high dimensional RNA-seq and picked the only the ideal features by utilizing BPSO-DT. Further enhanced RNA-seq was transformed to 2D images. Stage 2 dealt with augmentation, which increased the dataset of 2086 samples to 5 times greater. Augmentation methods were picked according to achieving the least impact on deploying the features of the images. Augmentation phase, helped in overcoming the over-fitting issue and trained the model with the aim to accomplish better accuracy. Eventually, stage 3 dealt with Deep CNN architecture. 2 CL for feature extraction and 2 FCL was employed for classifying five different types of cancer. From the outcome, accuracy obtained was 96% [61].

Though, traditional DL based methods deliver decent performance, accuracy in terms of prediction of disease is lacking. Therefore, transformer based DL methods were employed for prediction the disease more accurately and effectively.

Transformer based architecture leveraged SAM which encoded high throughput GE and learnt demonstration which were computationally complicated and affluent. Nevertheless, when compared to different datasets for applications of NLP, GE comprised of several 100 of

1000s gene from the restricted amount of interpretations, which made it hard to proficiently train the transformers for different applications of bioinformatics. Therefore, end to end DL method utilized. A gene transformer (GT) was used, as it addressed the difficulty of high dimensional GE with a multi-head SA component by detecting the appropriate biomarkers across the various multiple cancer subtypes without needing feature selection as a condition for the existing algorithms. The recommended paper enhanced the overall performance of the model. The outcome of the classification aided that gene transformer could be considered as an effective method for classifying the subtypes of the cancer, which indicated that any enhancement in DL model in computational biology could be reflected. GT architecture was motivated from transformer encoder architecture which used the idea of multi-head SAM with 1D CL as a hybrid architecture with the aim to evaluate the HD GE datasets and explored whether the demonstration leant from the AM which achieved better than the many prevailing methods. Gene transformer was considered as an end-to-end method which prioritized features during the testing phase and outperformed the prevailing techniques for both binary and multiclass issues. Gene transformer technique, was considered as a preliminary challenge in order to investigate the attention mechanism which was employed to predict the subtypes of lung cancer. In gene expression dataset, the sample of patients were oriented in 1D array. Hence, conventional CNN model was not considered as a viable technique for classification of tumor. Hence, suggested paper employed improvised version of 2D CNN model, which was termed as 1D-CNN model [16].

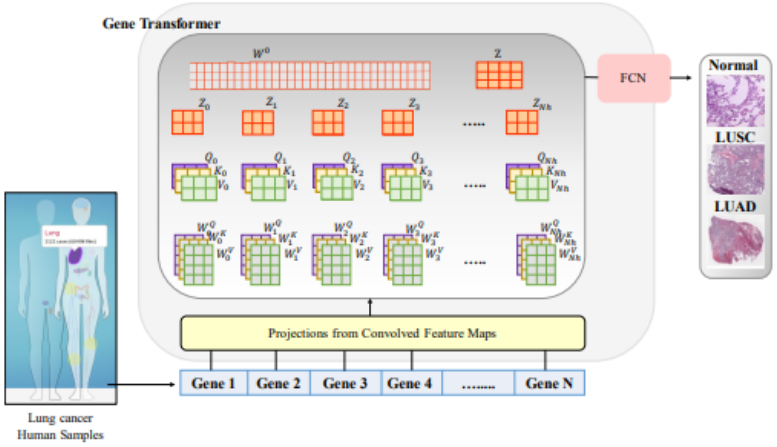

Figure 3. Graphical representations of the end to end DL for lung cancer subtype classification [16].

Recommended paper, employed miRe2e, which was the end – end DL technique for pre-miRNA prediction. miR2e approach was based on transformers, which is a neural architecture that employed AM (Attention Mechanism) with the aim to deduce the global dependencies between various inputs and outputs. It possessed the ability of getting the raw genome-wide data as input without any need of feature engineering or pre-processing process. This model aided in delivering more enhanced outcome [62].

CNN + Transformer hybrid architecture for GE level as well as epigenetic feature prediction was employed in enformer46, as enformer 46 was considered to be the first to do so. Thus, the model possess the potential to capture improved biological phenomenon, which includes enhancers that aided in controlling the promoters regardless of distance between the huge sequences of DNA. Therefore, it led to upsurge of significant performance in cell-type specific

gene expression prediction correlations [63]. Similarly, a DL technique has been employed for single cell category prediction by amalgamation of CNN with transformer with the aim to extract the powerful features. I-TM is used for extraction of features of the gene expression matric. Further, features were extracted using CNN model [64].

The unique characteristics of the gene expression aids the interpretable DL models which was personalized for transcriptomics study using different DL models. However, there are few encounters in evolving the DL techniques for modelling the expression of genes, hence suggested study employed T-GEM model for modelling the interactions of gene-gene and confirmed the usefulness for GE based prediction of the cancer associated phenotypes which included prediction of types of cancer and immune cell type classification. Hence, the mechanism of T-GEM was examined carefully and demonstrated that higher layer focused on phenotype related genes, whereas the first layer focused on broader attention. Therefore, the ability of T-GEM on the prediction of cancer type was demonstrated by utilizing TGCA data and the immune cell type of identified using scRNA-seq data [13].

The objective of the study was to classify the subtypes of the tumor and determining the significant gene in each subtype. This aided in evolving an amalgamated technique with the aim to detect the early-stage of cancer or distinguish the subtypes of the tumor. Over the past decade, both DL and ML techniques have been employed for classifying the subtypes of cancer from the GE dataset. However, these techniques were theoretically biased toward the detection of biomarkers of cancer. Therefore, DeepGene transformer, which was considered as an end- to-end DL method, addressed the complications of high dimensional gene expression with the multi-head self-attention module (MHSA). This module aided by detecting the relevant biomarkers across multiple-subtypes of cancer without any need of feature selection as a precondition for the classification of the algorithm [65].

In the last few years, different traditional ML and DL algorithms were used for disease classification using gene expressions. However, in recent days, VTN (Vision Transformer Networks) have delivered favorable performance in countless fields owing to prevailing AM (Attention Mechanism) which provided restored perception into data features. Still the network models have not been reconnoitered for analysis of gene expression. Therefore, suggested paper performed dimensionality reduction using a stacked auto-encoder which was followed by using an improved deep insight (I-DI) algorithm which converted the data into image format. The recommended study performed three techniques which comprised of 3 stages, which includes pre-processing, conversion of the data to image and classification. Further, classification model was built by feeding the data to the vision transformer. I-DI worked on principle of t-SNE (t-distributed Stochastic Neighbor Embedding). Further, all relevant genes, which was picked by the stacked auto-encoder were preserved by employing channel expansion algorithm, with the aim to serve the vision of the transformer in image format. Hence, the existing algorithm outperformed the prevailing algorithms in terms of selection of gene and cancer classification approaches. The dataset which was incorporated was GeneViT dataset, which was available for both binary and multiclass classification [66].

5.1.2. Stomach cancer

Stomach cancer which was associated of genes are associated for treatment of cancer and aided in serving as significant biological markers in order to evaluate diagnosis or prognosis, but also helped in giving reaction for the treatment. Consequently, it is significant to identify the candidate genes (CG) and examine the data for stomach cancer. Hence, suggested study amalgamated RNA-Seq gene expression data for stomach cancer which was acquired from TCGA

(An Open DB with Clinical Data), which examined for CG and examined these genes using a DL algorithm. In specific, candidate genes were chosen, as they could be distinguished between the patients between the cancer and healthy patients by PCA. A classification was attained by PCA method and the accuracy of the value was predicted, however, the accuracy was predicted to be lower than the other methods [67].

Auto-encoder was used for reducing the count of features. Further, recommended paper focused on feature reconstruction and aided in eliminating the noise within the data and the features with 0 variance across samples. This model facilitated in feature extraction with highest variances, which influenced the probabilities of survival, moreover, it estimated the probability of the survival for each patient by applying cox regression and random survival forest. Then, GPU (Graphical processing unit) was applied for speeding up the process. Eventually, the model was assessed and compared with various prevailing models on 3 different datasets on the basis of calibration curve, concordance index and run time. Eventually, GO molecular functions and biological pathways were evaluated for important genes [68].

Author has stated that, cancer have been diagnosed depending on the expression of genes through DNA microarray technology. The critical tactic for detecting the tumor are delivered by clustering the genes expression data. Nonetheless, the expression of gene data is accustomed as high-proportionate and substantial scale. Nearly, a tiny quantity of genes are required for diagnosis of cancer, even though the search space could be high. In the course of the process of classification, repetition and proportion of the gene expression data was minimized by feature selection. Hence, an organized and dominant feature selection algorithm was employed for fastening the learning function of the classifiers and aided in regulating the classification correctness [69]. Similarly, Experts aided in discovering the likelihood of the diseases, with the help of large amount of gene expression data, which has been identified via DNA microarray technique. This technique aided in providing optimistic outcome during the last few years. Though, there are few challenges which needs to be addressed. The dataset employed in the study was gene dataset and the model which was implemented in the recommended paper was DNN algorithm. From, the experimental outcome, it was identified that DNN algorithm delivered accuracy rate of 72.5%, when compared to the prevailing algorithms [1].

5.2. Drug discovery

The detection and enlargement of drug is a long, expensive, daunting and ineffective procedure. Regardless of advancements in technology in last 2 years, there has been some issues related to discovery of drugs with gene expression. Therefore, in the past data, it has been identified that, advancements in the information technology and growth of automation has led to production and collection of huge quantity of compound activity and biomedical data. Therefore, improvement in the realm of DL has should be employed for discovery of drug. Hence, recommended paper utilized a deep neural model termed DTSyn (Dual Transformer model for drug pair Synergy prediction) based on multi-head attention mechanism to identify different drug combinations. Hence, the paper focused on designing a fine-granularity transformer for capturing chemical substructure-gene and gene-gene associations and a coarse-granularity transformer for extracting chemical-chemical and chemical-cell line interactions. DTSyn can extract interactions among chemicals and cell lines, which may represent the mechanisms of drug action. Different existing methods only consider extracting chemical-cell line associations from one granularity, neglecting other dimensional interactions. To address the

aforementioned problems, recommended paper implemented a dual-transformer based deep neural network named DTSyn (Dual-Transformer neural network predicting Synergistic pairs) for predicting potential drug synergies. From the experimental outcome, it was identified that DTSyn utilizing dual-transformers has the great potential in identifying novel synergistic drug pairs and also providing possible interpretability in mechanisms of drug actions [70].

Likewise, DeepCE utilized transformer, which can learn about biological relations between chemical substructures and genes, to predict drug-induced expression profiles. Moreover, DeepCE utilizes a graph neural network and multihead attention mechanism to model chemical substructure–gene and gene–gene associations—for predicting the differential gene expression profile perturbed by de novo chemicals. Furthermore, data augmentation method that extracts useful information from unreliable experiments in the dataset has been used in the suggested paper [71].

Correspondingly, suggested study has employed 2 discrete datasets, which is induced for gene expression profiles by from open TG-GATEs and existence of information from FAERS DB, which could be applied for various ADRs. It incorporated data cleaning and data filtering along with hyper parameter tuning and feature selection. By employing DNN, 14 predictive approaches were built with a validation accuracy of 89.4%, which displayed the predicted ADRs reliably and steadily. From the findings it was assumed that recommended model was beneficial for detection of drug discovery [72].

The prediction of sensitivity and selection of drug responsive biomarker was considered as one of the key steps in discovery of drug. Different computational approaches have been established to assist the purpose, comprising various DNN techniques. Nevertheless, the modular relations among the genomic features was hugely unnoticed in the models. Therefore, with the aim to overcome the limitations, the role of the gene co-expression network on prediction of drug sensitivity was explored in the suggested study. Therefore, network based approach was first detected for representative feature for prediction of drug response by employing the gene co-expression network. Further, 2 Graph based NN techniques were employed and both these models were integrated for gene network information straight into NN for prediction of outcome. Further, the model was compared with other existing techniques such as SVR, RF, Elastic Net, and Partial Least Square Regression and other DNN approaches for prediction of the drug sensitivity. The relation between the genomic features are considered to be robust and steady when compared to correlation between each individual genomic features and drug response in more dimension and low size of sample in genomic datasets [73].

Suggested study has employed DeepTTA model which utilized transformer for representation of drug and learning and MNN for transcriptomic data prediction of anti-cancer drug responses. DeepTTA model transcriptomic GE data and substructures of drugs for the prediction of drug response [74]. From the experimental outcome, it was identified that, DeepTTA model achieved better performance in terms of all RMSE, PCCE (Pearson correlation co-efficient) and SRCC (Spearman's Rank Correlation Co-Efficient). Therefore, it was identified that, DeepTTA model was considered as an effective method for cancer drug design.

Similarly, response of drug was identified by integrating the DL graph regularized matrix factorization (Deep GRMF) along with graph models, NN, matrix factorization method with the aim to apply the diverse information from chemical structure of drug. Control of cellular signaling system and cancer cell states to forecast the response to drug. DeepGRMF learnt the

embedding's of the drugs so that the sharing of drug edifices as well as the mechanism of the actions which are correspondingly related in the embedding the space. Prediction performance of the model was enhanced by evaluating the DeepGRME model. Further, the model was effective for predicting the chemotherapy for lung cancer patients. The dataset implemented was TCGA dataset. Moreover, CCLE and GDSC dataset was retained as these datasets contained GE data of 2,758 genes in 954 and 477 cell lines distinctly. The DeepGRMF approach only utilized gene expression profiling data and additionally incorporating the genomic variations and epigenetic information, which will possibly enhance the performance of the model. The drug sensitivity forecasting was considered to be not as good as cell line prediction [75]. Table-4 shows the drug discovery employed using different architecture.

Table 4. Drug discovery using DL with Transformers

Year	Model	Architecture	Findings
2019	Multi Head Attention Molecular Transformer (BioNavi-NP) [76]	Transformer	89%
2021	DNN [72]	N/A	89.4%
2020	Molecular transformer [77]	Transformer	N/A
2022	SCROP [78]	Transformer	N/A
2022	DeepTTA [74]	Transformer	N/A
N/A – Not Available			

6. COMPARATIVE ANALYSIS

Various DL using transformers are used for prediction diseases based on their type, dataset, process are tabulated in the Table 5.

Table 5. Comparative analysis

References	Year	Model	Type	Dataset	Process
[16]	2021	Gene Transformer	Transformer	The Cancer Genomic Atlas (TCGA)	A gene transformer was used, as it addressed the difficulty of high dimensional gene expression with a multi-head SA module by detecting the appropriate biomarkers across the various several subtypes of cancer without needing feature selection as a pre-requisite for the existing classification algorithms. The outcome of the classification aided that gene transformer could be considered as an effective method for classifying the subtypes of the cancer, which indicated that any enhancement in DL model
[62]	2022	miRe2e	Transformer	Gene expression	miRe2e, which was the end – end DL technique for pre-miRNA prediction. miR2e approach was based on transformers, which is a neural architecture that employed AM (Attention Mechanism) with the aim to deduce the global dependencies between various inputs and outputs

[25]	2022	DeepTTA	Transformer	Gene expression	DeepTTA model utilized transformer for representation of drug and learning and MNN for transcriptomic data prediction of anti-cancer drug responses. DeepTTA model transcriptomic GE data and substructures of drugs for the prediction of drug response. From the experimental outcome, it was identified that, DeepTTA model achieved better performance in terms of all RMSE, PCCE (Pearson correlation co-efficient) and SRCC (Spearman's Rank Correlation Co-Efficient).
[66]	2023	Improved deep insight (I-DI)	Transformer	10 Benchmark Datasets	I-DI worked on principle of t-SNE (t-distributed stochastic neighbor embedding). Model performed three techniques which comprised of 3 stages, which includes preprocessing, conversion of the data to image and classification.
[65]	2023	DeepGene	Transformer	Gene Expression Dataset	DeepGene transformer, which was considered as an end- to-end DL method, addressed the complications of high dimensional gene expression with the multi-head self-attention module (MHSA). This module aided by detecting the relevant biomarkers across multiple-subtypes of cancer without any need of feature selection as a precondition for the classification of the algorithm

From the table, it can be interpreted that, transformers are used more from the year 2021. Different existing studies from 2021 has started using transformers in their model along with different datasets. Various models such as gene transformer, miRe2e, DeepTTA, Improved deep insight (I-DI), DeepGene models are implemented by various existing studies.

7. CRITICAL ANALYSIS

An analysis is carried out with respect to various aspects used by different traditional works prediction of diseases. 13 paper are considered for analysis. Applications of transformers for gene expression dataset have been mentioned in the Fig. 4. The obtained results are graphically presented in Fig. 4 have been considered.

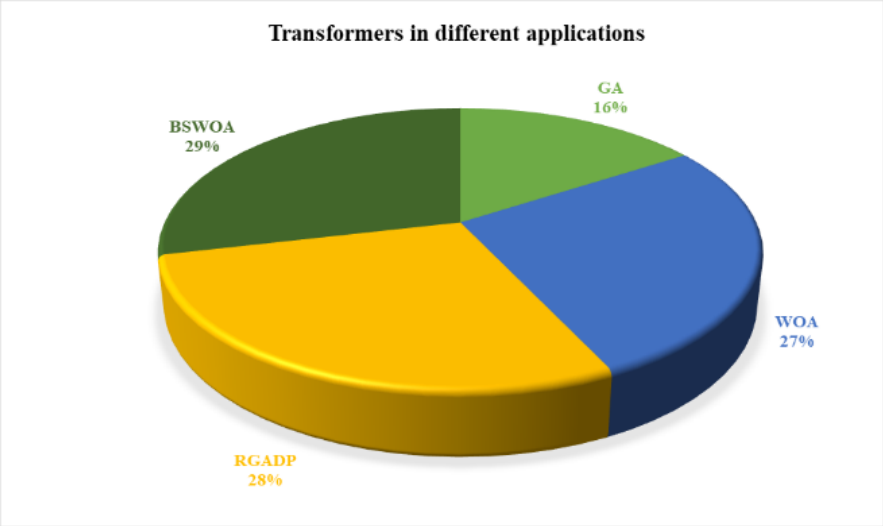

Figure 4. Transformers in different application.

Figure 4 shows the classification of different papers employed based on the topic. 46% shows the paper which dealt with prediction of lung cancer, 15% of papers dealt with stomach cancer and rest of the 8% papers dealt with prediction of other diseases and finally, rest of the, 31% of the papers dealt with drug discovery. Figure 5 shows the analysis of papers which involves DL models with transformers and without transformers.

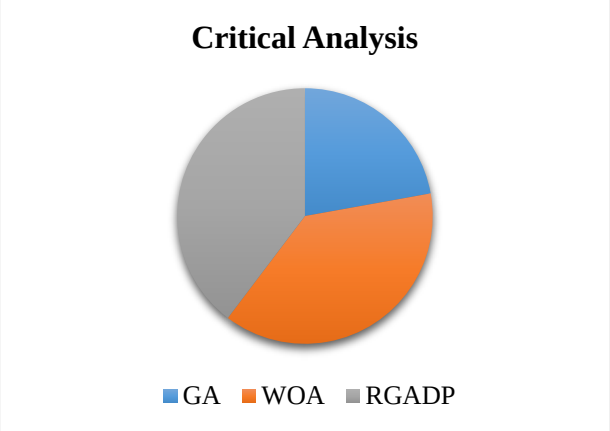

Figure 5. DL methods with transformers and without transformers.

8. FUTURE RECOMMENDATIONS

Various gaps identified through analysis of the existing works are listed below.

- Future work of the recommended paper focused on employing CNN for cancer multi-class classification [79].
- Future work of the suggested study focused on analyzing the function of the candidate genes and correlations in the expression depending on DL and incorporation of big data has the possibility to assist the detection of stomach cancer in early stage along with the development of treatment techniques and approaches with the aim to predict the prognosis [67].

- Though, ML has the potential to deliver satisfactory accuracy there are some of the issues like inability of the ML algorithms to employ unstructured data has restricted the utility of ML algorithms for classification tasks. Therefore, DL algorithms are preferred over ML, as they possess the ability to comprehend complex data.

Though there are various NN in DL, Transformers is utilized very less, therefore as a part of future work, disease can be predicted by employing various DL algorithm along with transformers, with the aim to predict accurate value for disease prediction, as Transformers models have demonstrated increased robustness and sturdiness more than models which delivered competitive performance on standards with different formats of data. One of the advantages of employing Transformer is that, the self-attention mechanism permits to use contextual information for any location in the input sequence and aids in capturing long range dependencies when compared to existing models and further, it allows higher parallelization when compared to RNN model.

9. CONCLUSION

The present study reviewed different experimenters work using DL for prediction of disease. Different DL algorithms and DL algorithms using transformers were examined in the present study. The corresponding features of related methods based on DL algorithms using transformers along with related datasets were presented. Based on DL algorithms and DL with transformers, prediction of diseases were found effectively. These transformers aids in delivering great accuracy and are efficient in terms of prediction of diseases. Since transformers possess different advantages such as increased capacity of the model, which allows to capture the complex relationships in sequential data, its ability to process the sequences in parallel, which makes it efficient and faster than conventional models. However, despite having several advantages, there have been limited studies in terms of using it prediction of diseases. Further, different applications like cancer prediction, drug discovery are mentioned in the present study. In addition, it has been observed that, incorporation of transformers in DL algorithms were implemented more from the year 2021, still there are few complications which needs to be overcome for prediction of diseases. Therefore, DL using transformers along with gene datasets will be helps in predicting the diseases. Additionally, this study would aids the researchers for their innovations in this area for enhancing the effectiveness and efficiency in prediction of disease.

CONFLICT OF INTEREST

All authors report that they do not have any conflicts of interest.

REFERENCES

1. V. Chandrasekar, V. Sureshkumar, T.S. Kumar & S. Shanmugapriya. Disease prediction based on micro array classification using deep learning techniques (Retracted). *Microprocessors and Microsystems*, **77**, 103189 (2020). Doi: https://doi.org/10.1016/j.micpro.2020.103189
2. Y. Yang, M.Z. Hossain, E.A. Stone & S. Rahman. Exemplar guided deep neural network for spatial transcriptomics analysis of gene expression prediction. 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). URL: https://ieeexplore.ieee.org/document/10030793

3. G. Lopez-Garcia, J.M. Jerez, L. Franco & F.J. Veredas. Transfer learning with convolutional neural networks for cancer survival prediction using gene-expression data. *PLoS One*, **15**(3), e0230536 (2020). Doi: https://doi.org/10.1371/journal.pone.0230536
4. Uzma, F. Al-Obeidat, A. Tubaishat, B. Shah & Z. Halim. Gene encoder: a feature selection technique through unsupervised deep learning-based clustering for large gene expression data. *Neural Computing and Applications*, **34**, 8309–8331 (2022). Doi: https://doi.org/10.1007/s00521-020-05101-4
5. S. Chowdhury. *Cell type classification via deep learning on single-cell gene expression data.* Ph. D. thesis. Prairie View A&M University, Prairie View (TX), 2023; 147 p. URL: https://digitalcommons.pvamu.edu/pvamu-dissertations/15/
6. M.B. Schabath & M.L. Cote. Cancer progress and priorities: Lung cancer. *Cancer Epidemiol. Biomarkers Prev.*, **28**(10), 1563–1579 (2019). Doi: https://doi.org/10.1158/1055-9965.epi-19-0221
7. J. Devlin, M.-W. Chang, K. Lee & K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT 2019, pp. 4171–4186. Minneapolis, Minnesota, June 2 - June 7, 2019. URL: https://aclanthology.org/N19-1423.pdf
8. Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q.V. Le & R. Salakhutdinov. Transformer-XL: Attentive language models beyond a fixed-length context. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 2978–2988. Florence, Italy, July 28 - August 2, 2019. URL: https://aclanthology.org/P19-1285.pdf
9. A. Kolesnikov, A. Dosovitskiy, D. Weissenborn, G. Heigold, J. Uszkoreit, L. Beyer, *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. Preprint, 2021. URL: https://arxiv.org/pdf/2010.11929v1
10. T. Brown, B. Mann, N. Ryder, M. Subbiah, J.D. Kaplan, P. Dhariwal, *et al.* Language models are few-shot learners. Preprint, 2020; 75 p. URL: https://arxiv.org/pdf/2005.14165
11. Z. Yang, A. Mitra, W. Liu, D. Berlowitz & H. Yu. TransformEHR: transformer-based encoder-decoder generative model to enhance prediction of disease outcomes using electronic health records. *Nat. Commun.*, **14**, 7857 (2023). Doi: https://doi.org/10.1038/s41467-023-43715-z
12. Q. Xu, L. Zhu, T. Dai & C. Yan. Aspect-based sentiment classification with multi-attention network. *Neurocomputing*, **388**, 135–143 (2020). Doi: https://doi.org/10.1016/j.neucom.2020.01.024
13. T.-H. Zhang, M. M. Hasib, Y.-C. Chiu, Z.-F. Han, Y.-F. Jin, M. Flores, *et al.* Transformer for gene expression modeling (T-GEM): An interpretable deep learning model for gene expression-based phenotype predictions. *Cancers*, **14**(19), 4763 (2022). Doi: https://doi.org/10.3390/cancers14194763
14. M.A. Osseni, P. Tossou, F. Laviolette & J. Corbeil. MOT: a Multi-Omics Transformer for multiclass classification tumour types predictions. Proceedings of the 16th International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 3: BIOSTEC, pp. 252–261, 2023, Lisbon, Portugal. Doi: https://doi.org/10.5220/0011780100003414
15. Z. Lv, Y. Lin, R. Yan, Z. Yang, Y. Wang & F. Zhang. PG-TFNet: Transformer-based fusion network integrating pathological images and genomic data for cancer survival analysis. IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2021. URL: https://ieeexplore.ieee.org/document/9669445
16. A. Khan & B. Lee. DeepGene Transformer: Transformer for the gene expression-based classification of cancer subtypes. *Expert Systems with Applications*, **226**, 120047 (2023). Doi: https://doi.org/10.1016/j.eswa.2023.120047
17. J. Wensel, H. Ullah & A. Munir. ViT-ReT: Vision and recurrent transformer neural networks for human activity recognition in videos. *IEEE Access*, **11**, 72227–72249 2023. URL: https://ieeexplore.ieee.org/abstract/document/10177697
18. I.D. Luptáková, M. Kubovčík & J. Pospíchal. Wearable sensor-based human activity recognition with transformer model. *Sensors*, **22**(5), 1911 (2022). Doi: https://doi.org/10.3390/s22051911
19. D. Huang, P.Z. Kosentka & W. Liu. Synthetic biology approaches in regulation of targeted gene expression. *Current Opinion in Plant Biology*, **63**, 102036 (2021). Doi: https://doi.org/10.1016/j.pbi.2021.102036

20. A. Anna & G. Monika. Splicing mutations in human genetic disorders: examples, detection, and confirmation. *J. Appl. Genet.*, **59**(3), 253–268, (2018). Doi: https://doi.org/10.1007/s13353-018-0444-7
21. B.E. Slatko, A.F. Gardner & F.M. Ausubel. Overview of next-generation sequencing technologies. *Curr. Protoc. Mol. Biol.*, **122**, e59 (2018). Doi: https://doi.org/10.1002/cpmb.59
22. N. Briglia, A. Petrozza, F.A. Hoeberichts, N. Verhoef & G. Povero. Investigating the impact of biostimulants on the row crops corn and soybean using high-efficiency phenotyping and next generation sequencing. *Agronomy*, **9**(11), 761 (2019). Doi: https://doi.org/10.3390/agronomy9110761
23. T. Phan, E.J. Fay, Z. Lee, S. Aron, W.-S. Hu & R.A. Langlois. Segment-specific kinetics of mRNA, cRNA, and vRNA accumulation during influenza virus infection. *J. Virol.*, **95**(10), e02102-20 (2021). Doi: https://doi.org/10.1128/JVI.02102-20
24. G. Monaco, B. Lee, W. Xu, S. Mustafah, Y.Y. Hwang, C. Carré, *et al.* RNA-Seq signatures normalized by mRNA abundance allow absolute deconvolution of human immune cell types. *Cell Reports*, **26**(6), 1627–1640.E7 (2019). Doi: https://doi.org/10.1016/j.celrep.2019.01.041
25. M.F. Khan, T.M. Ghazal, R.A. Said, A. Fatima, S. Abbas, M. Khan, *et al.* An iomt-enabled smart healthcare model to monitor elderly people using machine learning technique. *Comput. Intell. Neurosci.*, **2021**, 487759 (2021). Doi: https://doi.org/10.1155/2021/2487759
26. S.B. Bhonde & J.R. Prasad. Deep learning techniques in cancer prediction using genomic profiles. 2021 6th International Conference for Convergence in Technology (I2CT), Maharashtra, India, 2021; pp. 1-9. Doi: https://doi.org/10.1109/i2ct51068.2021.9417985
27. L.N. Waylen, H.T. Nim, L.G. Martelotto & M. Ramialison. From whole-mount to single-cell spatial assessment of gene expression in 3D. *Commun. Biol.*, **3**, 602 (2020). Doi: https://doi.org/10.1038/s42003-020-01341-1
28. S. Kumar, K. Dhial & T. Bhati. DNA microarray: Gene expression and analysis technology. *The Science World: A Monthly e-Magazine*, **1**(06), 24–29 (2021). URL: https://www.researchgate.net/publication/361485777_DNA_Microarray_Gene_Expression_and_Analysis_Technology
29. Y. Wang, M. Mashock, Z. Tong, X. Mu, H. Chen, X. Zhou, *et al.* Changing technologies of RNA sequencing and their applications in clinical oncology. *Frontiers in Oncology*, **10**, 447 (2020). Doi: https://doi.org/10.3389/fonc.2020.00447
30. A. Negi, A. Shukla, A. Jaiswar, J. Shrinet & R.S. Jasrotia. Applications and challenges of microarray and RNA-sequencing. In: D.B. Singh & R.K. Pathak (editors). *Bioinformatics: Methods and Applications*. Academic Press, 2022; pp. 91–103. Doi: https://doi.org/10.1016/B978-0-323-89775-4.00016-X
31. S. Dang, M. Wen, S. Mumtaz, J. Li & C. Li. Enabling multicarrier relay selection by sensing fusion and cascaded ANN for intelligent vehicular communications. *IEEE Sensors J.*, **21**(14), 15614–15625 (2020). Doi: https://doi.org/10.1109/JSEN.2020.2986322
32. T. Matsubara, T. Ochiai, M. Hayashida, T. Akutsu & J.C. Nacher. Convolutional neural network approach to lung cancer classification integrating protein interaction network and gene expression profiles. *J. Bioinform. Comput. Biol.*, **17**(03), 1940007 (2019). Doi: https://doi.org/10.1142/S0219720019400079
33. S.H. Shah, M.J. Iqbal, I. Ahmad, S. Khan & J.J. Rodrigues. Optimized gene selection and classification of cancer from microarray gene expression data using deep learning. *Neural Comput. Applicat.*, 1-12 (2020). Doi: https://doi.org/10.1007/s00521-020-05367-8
34. H. Fang, J.-U. Lee, N.S. Moosavi & I. Gurevych. Transformers with learnable activation functions. Findings of the Association for Computational Linguistics: EACL 2023, pp. 2382–2398, May 2-6, 2023. Doi: https://doi.org/10.18653/v1/2023.findings-eacl.181
35. M. Tripathi. Analysis of convolutional neural network based image classification techniques. *Journal of Innovative Image Processing*, **3**, 100–117 (2021). Doi: https://doi.org/10.36548/jiip.2021.2.003
36. A.K. Tyagi & A. Abraham (editors). *Recurrent Neural Networks: Concepts and Applications*. CRC Press, Boca Raton (FL), 2022. Doi: https://doi.org/10.1201/9781003307822

37. D. Deshwal & P. Sangwan. A comprehensive study of deep neural networks for unsupervised deep learning. In: A.E. Hassanien, R. Bhatnagar & A. Darwish (editors). *Artificial intelligence for sustainable development: Theory, practice and future applications*. Springer Cham, 2021; pp. 101–126. Doi: https://doi.org/10.1007/978-3-030-51920-9_7
38. C. Qiao, L. Yang, Y. Shi, H. Fang & Y. Kang. Deep belief networks with self-adaptive sparsity. *Applied Intelligence*, **52**, 237–253 (2022). Doi: https://doi.org/10.1007/s10489-021-02361-y
39. X. Chen, T.-H. Li, Y. Zhao, C.-C. Wang & C.-C. Zhu. Deep-belief network for predicting potential miRNA-disease associations. *Briefings in Bioinformatics*, **22**(3), bbaa186 (2021). Doi: https://doi.org/10.1093/bib/bbaa186
40. S. Garg, D. Tsipras, P. S. Liang & G. Valiant. What can transformers learn in-context? A case study of simple function classes. 31 p. URL: https://arxiv.org/pdf/2208.01066
41. S. Mehta, X. Lu, W. Wu, D. Weaver, H. Hajishirzi, J.G. Elmore, *et al.* End-to-End diagnosis of breast biopsy images with transformers. *Medical Image Analysis*, **79**, 102466 (2022). Doi: https://doi.org/10.1016/j.media.2022.102466
42. O. Ahmed & A. Brifcani. Gene expression classification based on deep learning. 4th Scientific International Conference, Najaf, IRAQ (4th SICN-2019). URL: https://ieeexplore.ieee.org/document/9019357
43. K. KC, R. Li, F. Cui, Q. Yu & A.R. Haake. GNE: a deep learning framework for gene network inference by aggregating biological information. *BMC Systems Biology*, **13**, 38 (2019). Doi: https://doi.org/10.1186/s12918-019-0694-y
44. R. Xie, J. Wen, A. Quitadamo, J. Cheng & X. Shi. A deep auto-encoder model for gene expression prediction. *BMC Genomics*, **18**, 845 (2017). Doi: https://doi.org/10.1186/s12864-017-4226-0
45. K.G. Dizaji, X. Wang & H. Huang. Semi-supervised generative adversarial network for gene expression inference. KDD '18: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018; pp. 1435–1444. Doi: https://doi.org/10.1145/3219819.3220114
46. J. Clauwaert, G. Menschaert & W. Waegeman. Explainability in transformer models for functional genomics. *Briefings in Bioinformatics*, **22**(5), bbab060 (2021). Doi: https://doi.org/10.1093/bib/bbab060
47. L. Jiao, G. Wang, H. Dai, X. Li, S. Wang & T. Song. scTransSort: Transformers for intelligent annotation of cell types by gene embeddings. *Biomolecules*, **13**(4), 611 (2023). Doi: https://doi.org/10.3390/biom13040611
48. S. Jubair, J.R. Tucker, N. Henderson, C.W. Hiebert, A. Badea, M. Domaratzki, *et al.* "GPTransformer: A transformer-based deep learning method for predicting Fusarium related traits in barley. *Frontiers in Plant Science*, **12**, 761402 (2021). Doi: https://doi.org/10.3389/fpls.2021.761402
49. Y. Liu, G. Sun, Y. Qiu, L. Zhang, A. Chhatkuli & L.V. Gool. Transformer in convolutional neural networks. URL: https://arxiv.org/pdf/2106.03180v2
50. L. Gai, M. Xing, W. Chen, Y. Zhang & X. Qiao. Comparing CNN-based and transformer-based models for identifying lung cancer: which is more effective?. *Multimedia Tools and Applications*, **83**, 59253–59269 (2024). Doi: https://doi.org/10.1007/s11042-023-17644-4
51. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, *et al.* Attention is all you need. 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA; 11 p. URL: https://dl.acm.org/doi/pdf/10.5555/3295222.3295349
52. A. Rives, S. Goyal, J. Meier, D. Guo, M. Ott, C.L. Zitnick, *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences* (PNAS), **118**(15), e2016239118 (2021). Doi: https://doi.org/10.1073/pnas.2016239118
53. A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. Published as a conference paper at ICLR 2021; arXiv: 2010.11929v2; 22 p.. Doi: https://doi.org/10.48550/arXiv.2010.11929
54. Z. Wang, Y. Ma, Z. Liu & J. Tang. R-Transformer: Recurrent neural network enhanced transformer. arXiv: 1907.05572v1, 2019; 11 p. Doi: https://doi.org/10.48550/arXiv.1907.05572

55. L. Chen, X. Pan, Y.-H. Zhang, M. Liu, T. Huang & Y.-D. Cai. Classification of widely and rarely expressed genes with recurrent neural network. *Computational and Structural Biotechnology Journal*, **17**, 49–60 (2019). Doi: https://doi.org/10.1016/j.csbj.2018.12.002
56. S. Pang, Y. Zhuang, S. Qiao, F. Wang, S. Wang & Z. Lv. DCTGM: A novel dual-channel transformer graph model for miRNA-disease association prediction. *Cognitive Computation*, **16**, 2009–2018 (2024). Doi: https://doi.org/10.1007/s12559-022-10092-6
57. T. Khorshed, M.N. Moustafa & A. Rafea. Deep learning for multi-tissue cancer classification of gene expressions (GeneXNet). *IEEE Access*, **8**, 90615–90629 (2020). Doi: https://doi.org/10.1109/access.2020.2992907
58. S. Parvathavardhini & S. Manju. Cancer gene detection using neuro fuzzy classification algorithm. *Int. J. Sci. Res. Comput. Sci., Eng. Inf. Technol.*, **3**(3), 1223–1229 (2018). URL: https://ijsrcseit.com/paper/CSEIT1833478.pdf
59. F. Gao, W. Wang, M. Tan, L. Zhu, Y. Zhang, E. Fessler, *et al.* DeepCC: a novel deep learning-based framework for cancer molecular subtype classification. *Oncogenesis*, **8**, 44 (2019). Doi: https://doi.org/10.1038/s41389-019-0157-8
60. M. Bicakci, O. Ayyildiz, Z. Aydin, A. Basturk, S. Karacavus & B. Yilmaz. Metabolic imaging based sub-classification of lung cancer. *IEEE Access*, **8**, 218470–218476 (2020). Doi: https://doi.org/10.1109/access.2020.3040155
61. N.E.M. Khalifa, M.H.N. Taha, D.E. Ali, A. Slowik & A.E. Hassanien. Artificial intelligence technique for gene expression by tumor RNA-Seq data: a novel optimized deep learning approach. *IEEE Access*, **8**, 22874–22883 (2020). Doi: https://doi.org/10.1109/access.2020.2970210
62. J. Raad, L.A. Bugnon, D.H. Milone & G. Stegmayer. miRe2e: a full end-to-end deep model based on transformers for prediction of pre-miRNAs. *Bioinformatics*, **38**(5), 1191–1197 (2022). Doi: https://doi.org/10.1093/bioinformatics/btab823
63. Ž. Avsec, V. Agarwal, D. Visentin, J.R. Ledsam, A. Grabska-Barwinska, K.R. Taylor, *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, **18**, 1196–1203 (2021). Doi: https://doi.org/10.1038/s41592-021-01252-x
64. T. Song, H. Dai, S. Wang, G. Wang, X. Zhang, Y. Zhang, et al., "TransCluster: A cell-type identification method for single-cell RNA-Seq data using deep learning based on transformer. *Frontiers in Genetics*, **13**, 1038919 (2022). Doi: https://doi.org/10.3389/fgene.2022.1038919
65. A. Khan & B. Lee. DeepGene Transformer: Transformer for the gene expression-based classification of cancer subtypes. *Expert Systems with Applications*, **226**, 120047 (2023). Doi: https://doi.org/10.1016/j.eswa.2023.120047
66. M. Gokhale, S.K. Mohanty & A. Ojha. GeneViT: Gene vision transformer with improved DeepInsight for cancer classification. *Computers in Biology and Medicine*, **155**, 106643 (2023). Doi: https://doi.org/10.1016/j.compbiomed.2023.106643
67. J. Xu, P. Wu, Y. Chen, Q. Meng, H. Dawood & H. Dawood. A hierarchical integration deep flexible neural forest framework for cancer subtype classification by integrating multi-omics data. *BMC Bioinformatics*, **20**, 527 (2019). Doi: https://doi.org/10.1186/s12859-019-3116-7
68. H. Torkey, M. Atlam, N. El-Fishawy & H. Salem. A novel deep autoencoder based survival analysis approach for microarray dataset. *PeerJ Computer Science*, **7**, e492 (2021). Doi: https://doi.org/10.7717/peerj-cs.492
69. H. Lu, J. Chen, K. Yan, Q. Jin, Y. Xue & Z. Gao. A hybrid feature selection algorithm for gene expression data classification. *Neurocomputing*, **256**, 56–62 (2017). Doi: https://doi.org/10.1016/j.neucom.2016.07.080
70. J. Hu, J. Gao, X. Fang, Z. Liu, F. Wang, W. Huang, *et al.* DTSyn: a dual-transformer-based neural network to predict synergistic drug combinations. *Briefings in Bioinformatics*, **23**(5), bbac302 (2022). Doi: https://doi.org/10.1093/bib/bbac302

71. T.-H. Pham, Y. Qiu, J. Zeng, L. Xie & P. Zhang. A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to COVID-19 drug repurposing. *Nature Machine Intelligence*, **3**, 247–257 (2021). Doi: https://doi.org/10.1038/s42256-020-00285-9
72. A. Mohsen, L.P. Tripathi & K. Mizuguchi. Deep learning prediction of adverse drug reactions in drug discovery using open TG–GATEs and FAERS databases. *Front. Drug Discov.*, **1**, 768792 (2021). Doi: https://doi.org/10.3389/fddsv.2021.768792
73. K.T. Ahmed, S. Park, Q. Jiang, Y. Yeu, T. Hwang & W. Zhang. Network-based drug sensitivity prediction. *BMC Medical Genomics*, **13**, 193 (2020). Doi: https://doi.org/10.1186/s12920-020-00829-3
74. L. Jiang, C. Jiang, X. Yu, R. Fu, S. Jin & X. Liu. DeepTTA: a transformer-based model for predicting cancer drug response. *Briefings in Bioinformatics*, **23**(3), bbac100 (2022). Doi: https://doi.org/10.1093/bib/bbac100
75. S. Ren, Y. Tao, K. Yu, Y. Xue, R. Schwartz & X. Lu. *De novo* prediction of cell-drug sensitivities using deep learning-based graph regularized matrix factorization. *Biocomputing*, **27**, 278–289 (2022). Doi: https://doi.org/10.1142/9789811250477_0026
76. S. Zheng, T. Zeng, C. Li, B. Chen, C. W. Coley, Y. Yang, *et al.* Deep learning driven biosynthetic pathways navigation for natural products with BioNavi-NP. *Nature Communications*, **13**, 3342 (2022). Doi: https://doi.org/10.1038/s41467-022-30970-9
77. G. Pesciullesi, P. Schwaller, T. Laino & J.-L. Reymond. Transfer learning enables the molecular transformer to predict regio-and stereoselective reactions on carbohydrates. *Nature Communications*, **11**, 4874 (2020). Doi: https://doi.org/10.1038/s41467-020-18671-7
78. S. Zheng, J. Rao, Z. Zhang, J. Xu & Y. Yang. Predicting retrosynthetic reactions using self-corrected transformer neural networks. *Journal of Chemical Information and Modeling*, **60**(1), 47–55 (2019). Doi: https://doi.org/10.1021/acs.jcim.9b00949
79. M.K. Elbashir, M. Ezz, M. Mohammed & S.S. Saloum. Lightweight convolutional neural network for breast cancer classification using RNA-seq gene expression data. *IEEE Access*, **7**, 185338–185348 (2019). Doi: https://doi.org/10.1109/access.2019.2960722

HOW TO CITE THIS ARTICLE

I. Mercy & M. Chidambaram. Prediction of diseases using transformers for gene expression dataset. *Rev. Colomb. Cienc. Quim. Farm.*, **54**(3), 591–620 (2025). Doi: https://doi.org/10.15446/rcciquifa.v54n3.118814