ORIGINAL RESEARCH ARTICLE UHD JOURNAL OF SCIENCE AND TECHNOLOGY

Utilizing Machine Learning Techniques for Cancer Prediction and Classification based on Gene Expression Data

Mariwan Mahmood Hama Aziz, Sozan Abdullah Mahmood

Department of Computer, College of Science, University of Sulaimani, Sulaymaniyah 46001, Kurdistan, Iraq

ABSTRACT

Cancer classification through genetic evaluation has become a hot topic among researchers. It holds the promise of delivering systematic, precise, and scientifically backed diagnoses for different types of cancer. Lately, several studies have delved into cancer classification by leveraging data mining techniques, machine learning algorithms, and statistical methods to thoroughly analyze high-dimensional datasets. Detecting cancer early by examining gene expression data is vital for providing effective patient care. Each sample in the Gene dataset usually includes a range of features, each representing a specific gene. In this paper, we propose a unique approach that utilizes DistilBERT, a distilled version of the Bidirectional Encoder Representations from Transformers, for cancer classification and prediction. In addition, our model integrates a self-attention mechanism in the transformer layers to enhance the model's focus on key features and employs an embedding layer for dimensionality reduction, improving the processing of gene statistics, preventing overfitting, and boosting generalization. We utilized datasets from important resources: The gene expression omnibus, which provided microarray records of lung and ovarian cancers, and the cancer genome atlas (TCGA), which offered RNA-Seq facts encompassing multiple most cancer types (breast invasive carcinoma, kidney renal clear cell carcinoma, colon adenocarcinoma, lung adenocarcinoma, and prostate adenocarcinoma). Our approach established excessive accuracy across all datasets, showcasing big upgrades in overall model performance compared to present strategies within the subject. The results underscore the ability to leverage transformer-primarily based architectures for strong cancer-type prediction and classification. Our approach achieved and improved exceptional accuracy compared to previous studies, with DS1: 97.56% for lung cancer, DS2: 100% for ovarian cancer, and DS3: 99.504% for the TCGA dataset.

Index Terms: Cancer Classification, Gene Expression Data, RNA-Seq, DNA Microarray, Bidirectional Encoder Representations from Transformers Model, Machine Learning, Pan-cancer, The Cancer Genome Atlas, DistilBERT

1. INTRODUCTION

Deoxyribonucleic acid, or DNA, stores genetic information needed by all living things to create, function, and develop. DNA is generally regarded as the blueprint of all living organisms since its components encode all of the information required to sustain life. Cancer is a complicated disease that stems from genetic mutations and unusual patterns of gene expression. These molecular shifts can throw off the normal functioning of cells, resulting in unchecked cell growth and the formation of tumors. Thanks to recent breakthroughs in gene expression profiling technologies, researchers can now assess the activity of thousands of genes all at once, offering crucial insights into how we diagnose, classify, and predict cancer outcomes [2], [3]. Cancer has become one of the most fatal illnesses globally, with an anticipated 9.7 million deaths from 20 million new cancer diagnoses in

Access this article online

DOI:10.21928/uhdjst.v9n1y2025.pp135-148

E-ISSN: 2521-4217
P-ISSN: 2521-4209

Corresponding author's e-mail: Mariwan Mahmood Hama Aziz, Department of Computer, College of Science, University of Sulaimani, Sulaymaniyah 46001, Kurdistan, Iraq. E-mail: mariwan.hamaaziz@univsul.edu.iq

Received: 05-04-2025 Accepted: 26-04-2025 Published: 02-06-2025

2022, according to the World Health Organization. Cancer is caused by the unrestrained proliferation of some abnormal cells, which divide and spread to other cells, multiplying malignant child cells. Men's most frequent cancers include lung, prostate, colorectal, and stomach [4]. Over the past two decades, health informatics research has focused on a variety of topics, including bioinformatics, cheminformatics, cancer prediction, and others [5].

Gene expression is the method by which the knowledge stored in DNA is transformed into instructions for producing proteins or other substances. It starts with the transcription of DNA into messenger RNA, which is then translated into proteins. Gene expression analysis is used to analyze the order of genetic modifications occurring under specific conditions in tissue or a single cell [6], [7]. A new technique for studying the expression of several genes at once is microarray technology. It entails positioning thousands of sequences of genes on a glass slide known as a "gene chip" in specific locations. The gene chip comes into contact with a sample of DNA or RNA. Measured light is produced by complementary base pairing between the sample and the gene sequences on the chip. Genes expressed in the sample are identified by regions of the chip that emit light. Each row in a tabular representation of a microarray gene expression data set corresponds to a single gene, each column to a sample or time point, and each matrix entry represents the measured expression level of a specific gene in a sample [8]–[10]. By offering more normalized and less noisy data for classification and prediction, RNA-Seq is a novel and well-liked method for finding new transcripts and isoforms. Finding the genes that are differentially expressed in a body or identifying changes in genes at various levels is the primary purpose of transcriptome profiling. RNA sequencing allows for both identification and quantification in one location. RNA-Seq data are widely available from various databases that can be used for cancer prediction and classification [11].

Machine learning (ML) techniques have recently been utilized to analyze microarray datasets for the categorization of cancer. One useful method for diagnosing cancer is to use the gene expressions found in microarray datasets. Several feature selection techniques have been used to identify the most important properties of malignant microarray datasets to enhance the performance of these widely used ML algorithms [6]. Notably, several innovative algorithms have surfaced that have demonstrated encouraging outcomes across a range of fields [5]. A subfield of artificial intelligence called ML gives computers the ability to learn from training data, identify patterns in data, and make predictions on their own that get better over time without explicit programming. Numerous classification techniques were developed in the ML field, and many of them were applied to the categorization of cancer [12].

In this paper, we advocate a technique called DistilBERT, which is a distilled version of the Bidirectional Encoder Representations from Transformers (BERT) model that retains 97% of BERT's language understanding power while being lighter, faster, and smaller. DistilBERT was first presented by Hugging Face and is designed especially for tasks requiring less processing power [13]. The BERT version, added by Wu *et al.* (2024), is a groundbreaking deep learning model designed for herbal language processing that has 110 million parameters for the base version and 340 million parameters for the large version. Unlike conventional models that study textual content input in a unidirectional manner. It uses a transformer structure that reads the input text bi-directionally. This lets it recognize the context of a phrase based totally on both its left and proper environment, imparting deeper semantic knowledge [14]. BERT's transformer-primarily based architecture has been adapted for diverse fields past textual content processing, such as bioinformatics and computational biology. In those packages [15]–[17]. DistilBERT, such as the BERT model, uses the same structure but is compressed to reduce model size, holding most of BERT's overall performance with fewer parameters – approximately 60% of the size of BERT (66 parameters), making it quicker and more efficient, and providing quicker predictions with high performance [18].

Two forms of gene expression datasets from diverse sources are used in this study to evaluate the efficacy of the recommended method, and the selected data do not achieve the high results with the previous model. The gene expression omnibus (GEO) provides microarray datasets, which include samples of ovarian and lung cancer. The availability and dependability of these microarray datasets, which provide a photo of gene activity, have made them famous for being used in most cancer studies [11], [19]. The 2nd set of statistics is derived from the cancer genome atlas (TCGA), a comprehensive RNA-Seq dataset that consists of facts on numerous cancer types, including prostate adenocarcinoma (PRAD), lung adenocarcinoma (LUAD), colon adenocarcinoma (COAD), kidney renal clear cell carcinoma (KIRC), and breast invasive carcinoma (BRCA). We can very well evaluate the adaptability and efficacy of the DistilBERT model across various gene expression technologies by way of the utilization of each microarray and RNA-Seq information [20], [21].

The structure of this paper is prepared as follows: The Methods section offers a complete evaluation of the datasets and pre-processing strategies employed. The architecture and implementation phase into the version of DistilBERT for numerical input and its integration into the cancer category framework. Finally, we gift the experimental results, comparing our method with existing today's models, observed with the aid of a dialogue at the implications and capacity applications of these studies in customized oncology and medical decision aid.

2. LITERATURE REVIEW AND PROBLEM STATEMENT

This section reviews key research in cancer classification, highlighting the transition from traditional methods to innovative approaches of deep learning and optimization techniques. It showcases studies that utilize gene expression data and delves into how metaheuristic algorithms have been employed to enhance feature selection and boost model performance.

2.1. ML-Based Methods

Tabassum *et al.* (2024) [3]. Proposed an ensemble learning approach that uses a bagging-based multilayer perceptron's and mutual information for feature selection to classify cancer from high-dimensional gene expression data. The method was applied to different cancer types, demonstrating its effectiveness in handling high-dimensional data and achieving varying levels of accuracy across several datasets. In this study, AbdElNabi *et al.* (2020) [4]. Introduced an intelligent decision support system for cancer classification using gene expression data from breast and colon cancers. Their method combines information gain (IG) for initial feature selection, Grey Wolf Optimization for further dimensionality reduction, and a support vector machine (SVM) for classification. Applied to microarray datasets, the approach effectively handled high-dimensional data and achieved strong classification performance, demonstrating its stability and reliability in early cancer diagnosis. Other studies by Guyon *et al.* (2002) [22]. Integrating recursive feature elimination (RFE) with SVM. The RFE method, used for gene choice, finished with incredible accuracy, and SVM for cancer classification consisted of 98% on leukemia datasets. A recent study introduced a two-phase hybrid feature selection method by Ali and Saeed (2023) [6]. Combining filter techniques (IG, gain ratio, Chi-squared) with genetic algorithms (GA) to improve cancer classification. The approach was tested using SVM, Naive Bayes, k-nearest associates (KNN), Decision Tree, and random forest (RF) on microarray datasets for breast, lung, Central Nervous System, and brain cancers. The GA step further refined features selected by filters, enhancing overall classification performance, Wei *et al.* (2023) [23]. Emphasized the importance of feature extraction and selection in high-dimensional gene expression data. They applied methods such as methods like principal component analysis (PCA), IG, and GA were broadly followed. A look at applying PCA with numerous classifiers, which includes choice trees (DT), SVM, and RF, performed variable outcomes, emphasizing the importance of effective function extraction in optimizing model overall performance, Li *et al.* (2020) [24]. Carried out an extensive study on pan-cancer classification, utilizing TCGA RNA-seq gene expression data from 31 different tumor types. They employed ML techniques to pinpoint groups of distinguishing genes that could differentiate between these tumor types with an impressive accuracy of over 90%. The research also delved into sex-specific variations in gene expression, underscoring the promise of certain biomarkers for tumor diagnosis and tailored treatment approaches. In this approach, García-Díaz *et al.* (2022) [25]. Proposed unsupervised studying strategies have additionally been explored for multiclass cancer classification. A look at employing an extreme learning machine with a genetic grouping algorithm completed a median accuracy of 98.8% for breast, kidney, and prostate cancers, demonstrating the feasibility of unsupervised techniques for high-dimensional data. In addition, Chen (2022) [26]. Presented ML models, which include SVM, linear discriminant analysis (LDA), and KNN, have also been explored for multi-cancer datasets, consisting of brain, prostate, and colon cancers. These fashions did F-scores above 80% and furnished insights into feature screening techniques for dealing with high-dimensional gene expression data. In another study, gene choice strategies have additionally been tailored for cancer classification by AlShamlan and AlMazrua (2024) [5]. An examination leveraging Harris Hawks Optimization and KNN completed perfect typing for colon tumors and leukemia datasets. These effects spotlight the promise of biostimulator algorithms in identifying biologically applicable gene markers. Mukhopadhyay *et al.* (2023) [12]. Proposed discriminant analysis (LDA) combined with RF is explored for excessive-dimensional microarray gene expression facts. The study finished with accuracies of 96% for breast cancer, 98% for most colon cancers, and 99% for most prostate cancers, demonstrating the effectiveness of dimensionality discount strategies in improving category overall performance for multi-cancer datasets. Brought a bendy category framework for cancer gene expression profiles by Hijazi and Chan (2013) [20]. Utilizing ML models, such as DT, RF, and KNN, they have a look at implementing more than one characteristic

choice strategy (filter out, wrapper, and embedded) to datasets together with leukemia, colon, and prostate cancer, showcasing the adaptability of ML frameworks throughout extraordinary cancer kinds.

2.2. Deep Learning and Hybrid Approaches

Similarly, another study by Das *et al.* (2023) [27]. Use CNN, LSTM, and hybrid architectures, such as DCNN-GRU with enhanced chimp optimization algorithms to classify cancer using microarray data. The researchers tested these models on datasets that included various subtypes such as brain, breast, prostate, colon, and leukemia. These approaches leverage deep learning capacity to capture complicated styles, supplying sturdy consequences in gene expression-based cancer detection, Yaqoob *et al.* (2023) [28]. Proposed Recent research has furthermore delivered hybrid algorithms to beautify most cancer classes; integrated ML classifiers for breast cancer classification, such as KNN, SVM, and Naive Bayes, with the sine cosine and cuckoo search algorithm (SCACSA) brought about high performance in breast cancer types, outperforming traditional techniques. The study presents limitations that are important to consider. The SCACSA method relies on the quality and size of the dataset used for validation. On the other hand, Tarek *et al.* (2016) [29]. The KNN set of rules has also proven promise in most cancer predictions. Have a look at applied wrapper, filter out, and embedded feature choice methods with microarray datasets for leukemia, colon, and breast cancers, achieving accuracy rates of 99% and 100%, respectively, showcasing the adaptability of KNN across exclusive cancer datasets. Rukhsar *et al.* (2022) [2]. Introduced a deep-learning framework for classifying multiple types of cancer using RNA-Seq gene expression data. They took the complex, high-dimensional gene data and converted it into 2D images through processes, such as normalization and zero-padding. Then, they employed eight different deep learning algorithms, including CNN, to extract features and categorize samples from five distinct cancer types. Their experiments, which involved various data splits and k-fold cross-validation, showed that CNN outshone the other models in terms of classification performance, achieving a high accuracy of 97%, Mohammed *et al.* (2023) [11]. Implemented hybrid stacking ensembles, which have furthermore proven powerful. For instance, employing 1D-CNN and LASSO with TCGA datasets yielded accuracies of 99.54 % for full datasets and 98.62% for reduced datasets, demonstrating the performance of deep learning with dimensionality reduction strategies. Some studies by Sucharita *et al.* (2024) [19]. Have centered on enhancing cancer type classification through deep learning improvements. For example, a hybrid version combining exponential sigmoid-deep notion networks and ranking methods carried out accuracies of 85–95% throughout seven cancer kinds, including leukemia and ovarian cancers, illustrating the potential of deep belief networks in gene expression evaluation. Aburass *et al.* (2024) [30]. Introduced a hybrid ML model combining CNN, LSTM, and GRU architectures for gene mutation category execution, achieving 80.6% accuracy, and suggesting opportunities for additional optimization in hybrid frameworks. Despite these improvements, challenges persist in attaining regular generalization throughout datasets and addressing the computational complexity of high-dimensional records evaluation. In previous work by Thakur *et al.* (2024) [21]. Multi-cancer analysis has, moreover, benefited from advancements in ML. A comprehensive genomic pan-cancer category using TCGA datasets was completed with 90% accuracy through integrating GA, demonstrating the value of function choice in large-scale genomic statistics evaluation. Another effort mixed RNN-CNN architectures with bottleneck function extraction, attaining accuracies of 97.8% for breast cancer and 99.4% for prostate cancer. In this study, Surbhi Gupta *et al.* (2023) [10]. Posited deep studying strategies continue to be pivotal for various cancer types. Deep learning on RNA sequence datasets was examined for breast, lung, kidney, prostate, and colon cancers. Although unique accuracy values are no longer certain, these studies demonstrate the strong potential of deep learning architectures in managing complex datasets, reinforcing their relevance in modern-day cancer studies. An innovative graph convolutional network (GCN) was applied to TGCA datasets by Martínez Logreira (2020) [31]. Attaining approximately 52% accuracy for pan-cancer evaluation. Although the performance became modest, this observation highlighted the potential of graph-based procedures for shooting complex relationships in genomic records. Table 1 provides a precise view of the literature discussed above.

2.3. Limitations of Existing Work

Although there are significant advances in cancer classification using gene expression data, several recurring challenges continue to limit the effectiveness and scalability of existing methods. Key limitations identified in recent studies include.

- Lack of generalization: A lot of models are trained and fine-tuned on specific datasets, but they often skip validation on external or diverse datasets.
- Dataset dependency: When sample sizes are small or when there's a heavy reliance on just microarray or RNA-Seq data, it limits how well these models can apply to a wider range of cancer types.
- Computational cost: Methods that rely on optimization, such as GA and deep learning frameworks, tend to be

TABLE 1: Comparative review of literature

References	Model	Feature extraction	Dataset	Year
[3]	Multilayer perceptron's (MLPs)	Mutual information algorithm	Microarray	2024
[29]	k-nearest neighbors (KNN) algorithm	Wrappers, Filters, Embedded methods	Microarray	2016
[27]	CNN, LSTM, DCNN, GRU, PSCS-DL, CSSMO-DL	ECO algorithm	Microarray	2024
[4]	SVM	Information gain (IG)	Microarray	2020
[6]	SVM, NB, KNN, DT, RF	IG, information gain ratio, and Chi-squared	Microarray	2023
[2]	CNN	Deep learning (DL)	RNA-Seq data	2022
[25]	Extreme learning machine (ELM)	Grouping genetic algorithm (GGA)	RNA-Seq data	2020
[11]	1D-CNN	LASSO	(TCGA)	2022
[26]	SVM, LDA, or KNN	Feature screening	Gene expressions	2022
[22]	SVM with RFE	Recursive feature elimination (RFE)	Leukemia data	2002
[19]	Exponential sigmoid-deep belief network (ES-DBN)	Feature ranking (CM-CRO,)	Microarray Data	2024
[5]	(KNN), (SVM),	Harris hawks optimization (HHO)	Microarray Data	2024
[12]	Linear discriminant analysis (LDA) and (RF)	Linear discriminant analysis (LDA)	Microarray	2024
[30]	LSTM, LSTM, CNN, GRU	Not mention	Cancer Treatment dataset	2024
[28]	SVM, KNN, NB	(SCACSA)	Microarray Data	2024
[23]	DT, SVM, RF, NB, Neural network, KNN	Principal component analysis (PCA)	Microarray Data	2023
[20]	DT, SVM, RF, KNN, bagging,	Filter, wrapper, and embedded methods	Microarray Data	2013
[24]	KNN	(GA)	(TCGA)	2017
[21]	RNN-CNN	Sandwich stacked method based on VGG16 and VGG19 pre-trained models	Gene expression data	2023
[10]	Deep learning	Not mentioned	RNA sequence dataset	2022
[31]	Graph convolutional network (GCN)	Genetic Algorithms (GA)	The TGCA dataset	2020

ECO: Enhanced chimp optimization

resource-heavy, making them less ideal for real-time applications or environments with limited resources.

- Manual or static feature selection: Many studies stick to traditional feature selection techniques that need manual adjustments and don't adapt on the fly during training.
- Limited data pre-processing and hyperparameter tuning: Some methods fall short on having effective pre-processing steps or optimized hyperparameter choices, which can hurt their overall performance.
- Limited use of advanced models: There's a noticeable lack of exploration into transformer-based or graph-based neural networks in many studies, even though these could do a better job of capturing complex relationships between genes.

2.4. Problem Statement

Cancer diagnosis remains a critical challenge in healthcare, where early and accurate detection is essential to improving outcomes and reducing mortality. Traditional methods often fall short in handling the complexity of gene expression data, and many ML approaches struggle with generalizability, static feature selection, and dataset-specific tuning. To address these issues, this study introduces a DistilBERT-based model with a self-attention mechanism that dynamically identifies significant gene features during training. This approach enhances accuracy, reduces manual pre-processing, and offers a scalable solution for classifying multiple cancer types using high-dimensional gene expression data.

3. MATERIALS AND METHODS

The main steps in developing this research for cancer classification using gene expression include data collection, data pre-processing, gene selection using the self-attention mechanism, and finally classification using the DistilBERT model. Fig. 1 describes the processing steps of the proposed methodology; each step is briefly described next.

3.1. Data Collection

To determine the effectiveness of our DistilBERT model for cancer classification across different cancer types, we utilized publicly available gene expression datasets from two open sources platform, such as the GEO and TCGA. These data are not re-identifiable and have been released under a license that prohibits their use for commercial purposes only. They were used in a way that matches the requirements, that is, under a subscription based upon the terms and conditions established by both GEO and TCGA. The study does not need the formal approval of the Institutional Review

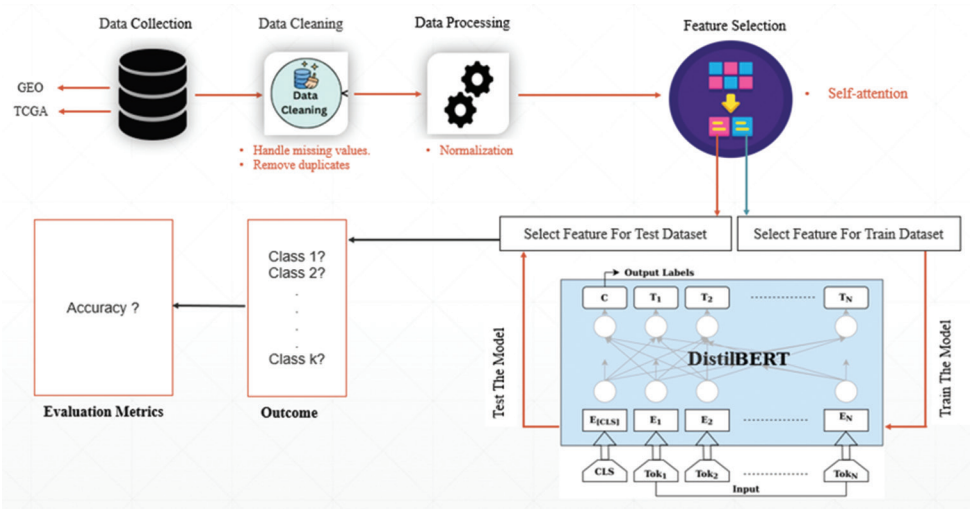

Fig. 1. Steps of the proposed methodology for cancer classification.

Board as it is public data and it does not contain any private information that has an identifiable person. We specifically selected datasets that had presented challenges to previous models, aiming to demonstrate the potential of our approach and achieve better results.

The GEO is managed by the National Center for Biotechnology Information. It is a repository for high-throughput gene expression and other functional genomics data. From GEO, we downloaded two high-dimensional microarray datasets: one for lung cancer and another for ovarian cancer [6] (as detailed in Table 2). The lung cancer dataset contains 203 instances with five classes and 12,600 features, or genes. The samples in the lung cancer dataset are classified as belonging to four classes of lung tumors: Small cell lung cancer (6 samples), adenocarcinoma (139 samples), normal lung (17 samples), squamous cell carcinoma (21 samples), and pulmonary carcinoid (20 samples) [2]. The ovarian cancer dataset includes 253 samples with two classes and 15,154 genes. The ovarian cancer dataset is labeled with a normal class (91 samples) and with cancer classes (162 samples) [3].

The second data source were TCGA, More than 20,000 primary cancers and matched normal samples from 33 different cancer types were molecularly characterized by the groundbreaking Cancer Genome Atlas (TCGA) program. Beginning in 2006, this collaborative effort between NCI and the National Human Genome Research Institute brought together scientists from various institutions and disciplines. In this source, we downloaded the RNA-Seq gene expression data from Pan-Cancer Atlas (https://portal.gdc.cancer.gov/) using the R statistical application version 3.6.3 by the TCGAbiolinks package [2], [11]. The dataset contains 801 instances or samples and 20,531 features or genes from the top five common cancer types, including BRCA, KIRC, COAD, LUAD, and PRAD [2], [11], [24]. Each sample has 20,532 gene sequences. The dataset's cancer classes are denoted by the following codes: 0, 1, 2, 3, and 4 for PRAD, LUAD, BRCA, KIRC, and COAD. Out of a total of 801 samples, the BRCA class has 300 samples, clear cell carcinoma (KIRC) has 146, LUAD has 141, COAD has 78, and PRAD class has 136 samples [21], [24]. As shown in Table 3, after downloading, we combine each type of cancer to make a unified, large-scale dataset for training and evaluating our model, aiming for a more generalized and accurate approach to cancer classification and prediction across multiple cancer types.

TABLE 2: Description of the high-dimensional microarray datasets used in this study

Dataset	No. of features	No. of instances	No. of classes
DS1: Lung cancer	12,600	203	5
DS2: Ovarian cancer	15,154	253	2

TABLE 3: Description of the DS3: Pan-cancer datasets used in this study

Dataset	No. of features	No. of instances
BRCA	20,532	300
KIRC	20,532	146
LUAD	20,532	141
COAD	20,532	78
PRAD	20,532	136

3.2. Data Processing

Before using a ML model, the selected datasets must be properly processed and processing raw gene expression data can be challenging due to its varied range. Several common procedures are taken during the pre-processing stage, including Data Cleaning. Normalization and feature selection.

3.2.1. Data cleaning

To ensure the quality and reliability of our datasets for effective model training, a data cleaning phase was performed. This involves identifying and dealing with different facets of data quality, including missing values, duplicates, inconsistencies, and outliers, which can lead to poor performance and interpretability of ML models. Removing missing values and duplicates is a very important step toward statistics cleaning to ensure the quality and abundance of the dataset, substituting missing values with statistical measures such as mean, median, or mode such as mean, median, or mode as shown in Tables 4 and 5. Similarly, duplicate records in a dataset can distort evaluation and version performance. Identifying and getting rid of duplicates guarantees statistics integrity and decreases redundancy [3], [27], [32].

3.2.2. Normalization

To ensure that all gene expression features contributed equally to the model training process, we applied normalization using the StandardScaler technique. The goal of normalization is to convert the values of numeric columns in the dataset to a common scale, which improves both the performance and accuracy of your model without distorting value ranges or losing any information [33]. We specifically employed StandardScaler, which centers the data around a mean of 0 and a standard deviation of 1. Figure 3 shows the data before normalization, and Figure 4 demonstrates the data after normalization. StandardScaler enhances version education balance by preventing features with larger scales from dominating other [34]. The scikit-learn (sklearn) library in Python includes the StandardScaler implementation. Fig. 2 is the form of the script we used when StandardScaler was implemented before data splitting was done.

```
scaler = StandardScaler ()
X_scaled = scaler.fit_transform(X)
```

Fig. 2. Standardscaler Implementation [1].

We use StandardScaler in normalization, and the equations (1), (2), and (3) represent the metaethical formula of standardization, mean, and standard deviation. Where X is the original value of the feature, N is the total number of values in the dataset, μ is the mean of the feature, and σ is the standard deviation of the features [35]–[37].

$$\text{X standardization} = \frac{x - \mu}{\sigma} \tag{1}$$

$$\text{Mean } \mu = \frac{1}{N}\sum_{i=1}^{N} Xi \tag{2}$$

$$\text{Standard Deviation } \sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Xi - \mu)2} \tag{3}$$

3.2.3. Feature selection

To identify the most relevant gene expression features for reliable cancer classification, we employed the inherent self-attention mechanism within the DistilBERT model for feature selection. This training process, learning their significance without explicit pre-processing, evaluates the relationships and dependencies among capabilities, assigning attention weights that reflect their importance in the context of the given project. Unlike traditional feature selection strategies, which necessitate either manual guidance or algorithmic

	AFFX-MurIL2_at	AFFX-MurIL10_at	AFFX-MurIL4_at	AFFX-MurFAS_at	AFFX-BioB-5_at	...	105_at
0	-18.600	10.54	0.010	19.440	-16.980	...	1.630
1	9.120	9.12	10.180	29.290	-4.680	...	10.180
2	-2.175	-2.21	-0.060	6.320	-1.775	...	1.745
3	-1.540	21.75	5.835	23.815	-24.785	...	10.355
4	-9.070	3.08	-1.980	17.260	-10.090	...	-10.090
..	...	...	...	...	...	...	...
198	35.140	106.16	52.280	65.340	27.790	...	48.200
199	-21.150	-31.20	-11.820	8.280	-24.740	...	-3.210
200	26.900	10.44	18.230	33.830	-11.220	...	6.970
201	23.800	29.14	31.800	65.610	4.240	...	26.470
202	-18.370	-1.03	-8.260	27.150	-23.430	...	-4.640

[203 rows x 12600 columns]

Fig. 3. Example of data before normalization.

pre-processing to pinpoint and eliminate irrelevant features, it dynamically learns which capabilities (genes) are maximally relevant for distinguishing between different cancer types. [38]–[42]. This concept is mathematically captured by what's known as scaled dot-product attention.

$$\text{Attention Score}(Q, K, V) = \text{softmax}\left(\frac{Qk^T}{\sqrt{d_k}}\right) V \quad (4)$$

Where Q, K, and V are the query, key, and value matrices derived from gene expression embeddings. By focusing on these dynamically identified key features, the model can potentially achieve better performance and generalization [39].

3.2.4. Data splitting

To train our DistilBERT model and see how well it performs on new, unseen data, we split each of our datasets into training and testing sets. Training data contained up to 80% of the overall dataset, allowing it to learn the patterns in the gene expression profiles associated with different cancer types, whereas test data represented up to 20%. As a result, the DistilBERT models were used to classify the cancer types.

```
[[-0.52270655  0.24721653 -0.02648141 ... -0.40856243 -0.9889013
   0.18736746]
 [ 0.82877923  0.18378861  0.50429813 ... -1.27053002  0.60904913
   1.18361247]
 [ 0.27809265 -0.3222947  -0.03013476 ... -0.48253143  0.41890506
   0.07492172]
 ...
 [ 1.69564132  0.24274977  0.92443336 ... -1.35286656  1.40594247
   0.01376094]
 [ 1.54450113  1.07803291  1.63266132 ...  1.90043074  0.55746503
   2.20064867]
 [-0.51149292 -0.26958699 -0.4580986  ...  0.93894528  0.82972729
  -1.08817267]]
```

Fig. 4. Example of data after normalization by Standardscaler.

4. PROPOSED CLASSIFICATION MODEL

Our proposed classification model uses a modified DistilBERT architecture to classify cancer types based on high-dimensional numerical gene expression datasets. This model is intended to efficiently process and evaluate input information to accurately forecast the cancer class and use our model with different types of cancer and achieve the highest accuracy.

4.1. DistilBERT

DistilBERT is a streamlined version of the BERT model, as shown in Fig. 5. It starts by taking in high-dimensional numerical inputs that represent gene expression levels. The first step is an embedding layer that uses a linear transformation (nn.Linear) to shrink the input dimensions down to 768, setting the stage for the next steps. To enhance training stability and improve generalization, a Group Normalization (Group Norm) layer is applied right after the embedding layer, ensuring that the input to the Transformer is appropriately normalized. After that, the model goes through a Transformer block made up of six stacked layers. Each of these layers features a self-attention mechanism, allowing the model to hone in on the most significant gene features by assigning varying weights across the input sequence. A special [CLS] token is added at the beginning, and it gets refined through these layers to serve

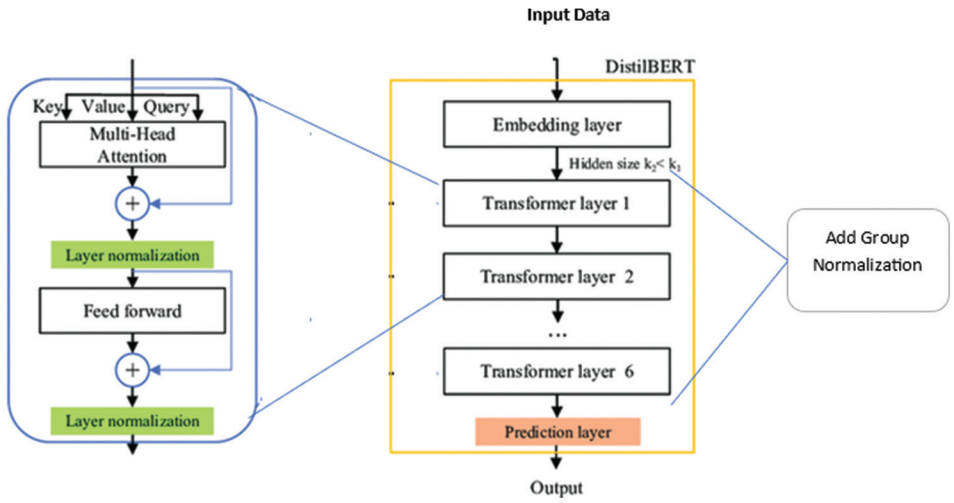

Fig. 5. Workflow of the DistilBERT model approach used as an ensemble classifier.

as a global summary of the entire sample. Following the self-attention process, a feed-forward network (FFN) – which is a fully connected network – adds non-linearity, enriching the feature representations and boosting the model's learning capabilities. Before reaching the final classification stage, a dropout layer with a rate of 0.2 is applied to the [CLS] token to help prevent overfitting and improve the model's generalization. Finally, a fully connected classification layer translates the 768-dimensional [CLS] representation into the output space, and the results are processed through a SoftMax function to generate the final class probabilities. One of the key strengths of the proposed study lies in its capability for generalization. By using a flexible transformer-based architecture, the version may be tailored to categorize extraordinary sorts of cancer with minimal changes. This generalizability becomes evident within the steady overall performance across various datasets examined throughout the study.

4.2. Training

During the training step, the model uses optimizers, such as AdamW and SGD with a learning rate of 2×10^{-5} and a weight decay of 1×10^{-3} and uses cross-entropy loss for classification, which allows the model to learn the relationships between the input features and the cancer classes.

4.3. Experimental Setup

The model we proposed was built using Python 3.11.5 and Visual Studio Code. We made use of several essential libraries. We preprocessed and normalized different types of cancer from two open sources. For cancer classification, we fine-tuned DistilBERT, a streamlined transformer model. This model features 6 transformer layers, 12 attention heads, and 768 hidden units and can handle a maximum input length of 512. All our experiments were conducted on a Windows 10 machine equipped with an Intel Core i7-6820HQ CPU, 16 GB of RAM, and an NVIDIA GeForce RTX 4060 GPU.

5. RESULTS ANALYSIS

The general efficacy of the proposed DistilBERT-based complete variant for most cancer classes was examined using different datasets, including lung cancer, ovarian cancer, and TCGA datasets. The results show the model's ability to efficiently handle high-dimensional gene expression data and achieve exceptional class accuracy across several datasets. The results are evaluated using accuracy performance metrics. In our experiments with the DistilBERT model, we explored various batch sizes and used the AdamW and SGD optimizers on all datasets. For the lung cancer dataset, our approach achieved an exceptional accuracy improvement of 97.56% with the AdamW optimizer and a batch size of 16, outperforming other batch sizes and SGD, making it the best choice for this task. Conversely, the SGD optimizer showed optimal performance with batch sizes of 32 and 128, as shown in Table 6. Fig. 6 displays the highest accuracy and loss attained by the model for lung cancer datasets, and Fig. 7 shows the confusion matrix model.

For the ovarian cancer dataset that contains 15,154 genes with two classes, our model gets 100% accuracy with the AdamW optimizer across all batch sizes; with the SGD, the results are represented in Table 7 with different batch sizes. Fig. 8 displays the highest accuracy and loss attained by the model for ovarian cancer datasets, and Fig. 9 shows the confusion matrix model.

For the TCGA dataset that includes five types of cancers and contains 20,532 genes for each type, we use DistilBERT with

TABLE 4: Example of data with miss values and duplicate rows

Sample ID	Feature1	Feature2	Feature3	Feature4
Sample_0	0.0	2.017209	3.265527	5.478487
Sample_1	0.0	0.592732	1.588421	7.586157
Sample_2	Nan	NaN	2.3271	6.881787
Sample_3	0.0	3.511759	4.327199	6.881787
Sample_4	0.0	3.511759	4.327199	6.881787

TABLE 5: Example of data after handle missing values with remove duplicate rows

Sample ID	Feature1	Feature2	Feature3	Feature4
Sample_0	0.0	2.017209	3.265527	5.478487
Sample_1	0.0	0.592732	1.588421	7.586157
Sample_2	0.0(Mean)	2.408865(Mean)	2.3271	6.881787
Sample_3	0.0	3.511759	4.327199	6.881787

TABLE 6: Using different batch sizes with AdamW and SGD optimizers for the lung cancer dataset

Batch size	Optimizer	Accuracy %
16	AdamW	0.9756
32		0.926
64		0.9268
128		0.9268
16	SGD	0.9024
32		0.92
64		0.878
128		0.92

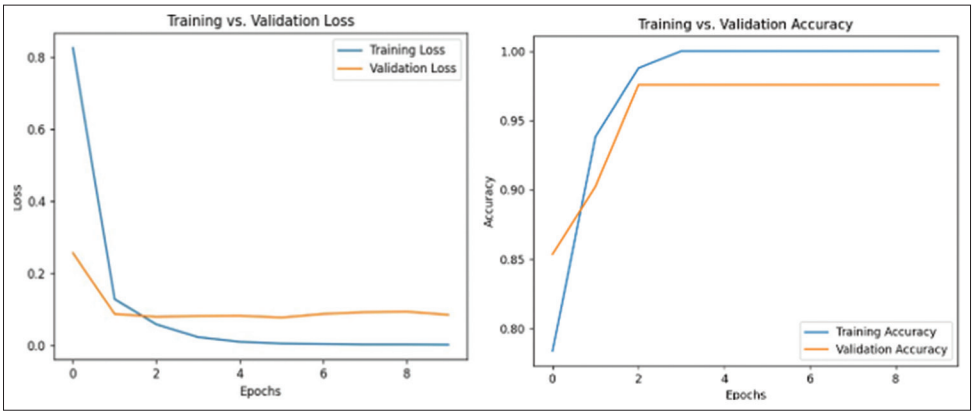

Fig. 6. Accuracy and loss for DistilBERT model for lung cancer dataset.

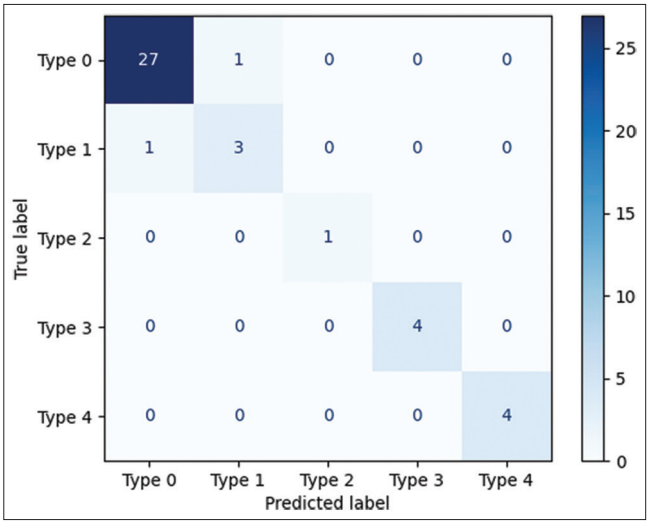

Fig. 7. Confusion matrix model for lung cancer dataset.

AdamW and SGD optimizers. Our model achieves 99.5% accuracy with the AdamW optimizer with a batch size of 16, as shown in Figs. 10 and 11 shows the confusion matrix. For the SGD optimizer, the result is presented in Table 8 with different batch sizes for both optimizers.

6. DISCUSSION

This paper proposed an approach by utilizing ML techniques for cancer prediction and classification based on gene expression data. The DistilBERT model was applied to classify gene expression datasets in bioinformatics because it can find complex non-linear relationships between the inputs and the outputs and is effective for large datasets. As discussed, it above indicates that DistilBERT model architectures could be used with two types of optimizers,

TABLE 7: Using different batch sizes with SGD for ovarian cancer

Batch size	Optimizer	Accuracy %
16	SGD	100
32		100
64		0.9804
128		0.9412

TABLE 8: Using different batch sizes with AdamW and SGD optimizer

Batch size	Optimizer	Accuracy %
16	AdamW	0.995
32		0.9937
64		0.9804
128		0.98
16	SGD	0.987
32		0.9813
64		0.9565
128		0.9255

SGD and AdamW, for cancer classification, as they show a confident result. The results indicate that model architectures performed well. The AdamW optimizer reached higher results across different batch sizes with model architecture, showing a better choice as its accuracy reached 97.56% with a batch size of 16, for the lung cancer dataset, while the accuracy dropped to (92.6%, 92%, and 92.68%) with batch sizes of 32, 64, and 128. For ovarian cancer the model performed well, and the accuracy achieved (100%) with the AdamW optimizer was the same with all different batch sizes. For the TCGA dataset, the accuracy reached (99.5%) with a batch size of 16, which is higher than the results for batch sizes of 32, 64, and 128 (99.37%, 98.04%, and 98%). On the other hand, the SGD optimizer produced the results in different batch size values, which are (16, 32, 64,

TABLE 9: Comparison of our proposed model with some existing works used in this field

Dataset	Existing work			Accuracy of our proposed model
	Method	Accuracy (%)	Year	
DS1: Lung Cancer	SVM, KNN, DT, RF [6]	94.09–97.04	2023	97.56%
	CNN [2]	97	2022	
	ES-DBN [19]	94.5454	2024	
	LDA and RF [12]	95	2024	
	RNN-CNN [21]	0.97	2023	
	SVM, RF, MLP, SMO [7]	93, 96, 86.6, 91	2022	
DS2: Ovarian Cancer	MLPs [3]	98	2024	100%
	1D-CNN [11]	98.62	2022	
	ES-DBN [19]	95.7746	2024	
	MI, GA, SVM [7]	80–98	2022	
DS3: TCGA Dataset	1D-CNN	98.62	2022	99.504%
	ELM [25]	98.81	2020	
	KNN [24]	90	2017	
	GCN [31]	52	2020	
	CNN [2]	97	2022	

SVM: Support vector machine, LDA: Linear discriminant analysis, MLPs: Multilayer perceptrons, KNN: k–Nearest Neighbors, ELM: Extreme learning machine, GCN: Graph convolutional network, ES-DBN: Exponential sigmoid-deep notion networks

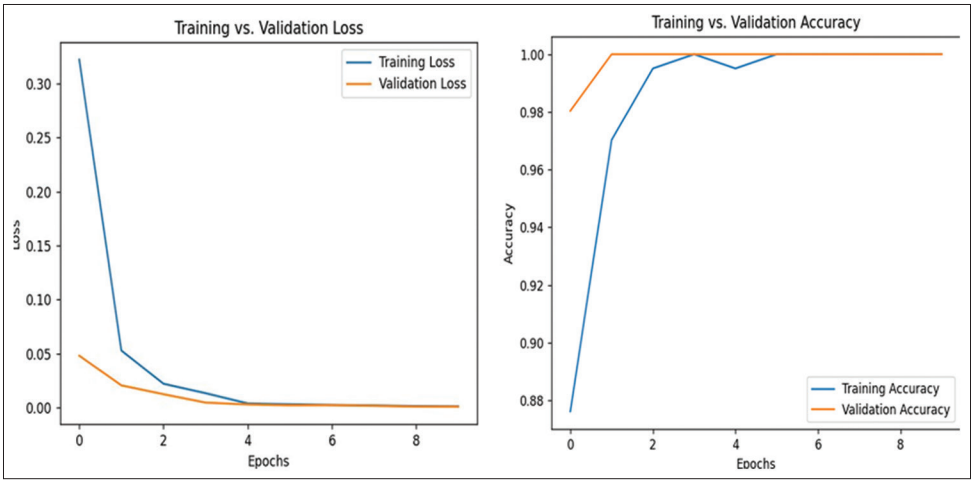

Fig. 8. Accuracy and loss for DistilBERT model for ovarian cancer dataset.

and 128). The accuracy was (90.24%, 92%, 87.8%, 92%) for lung cancer, but only in 16 and 32 did it reach the result of (100%), for ovarian cancer with model architecture. For the TCGA dataset, the accuracy is decreased with SGD for model architecture, as shown in Table 8. This suggests that batch sizes of 16 and 32 are the most effective for achieving optimal performance. This experiment discovered that the smaller batch size of 16 worked better because it allowed for improved generalization from the noisier updates, and the AdamW optimizer with batch 16 is more proper for cancer classification, as it got a high accuracy in almost all the tests compared to the SGD optimizer. The consistent 100% accuracy on the ovarian dataset with AdamW was unexpected, suggesting a highly separable gene expression

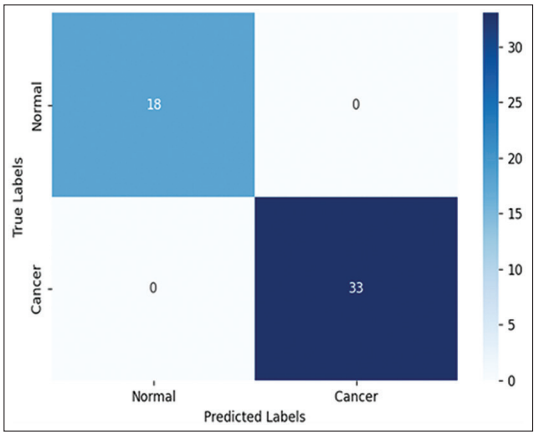

Fig. 9. Confusion matrix model for ovarian cancer dataset.

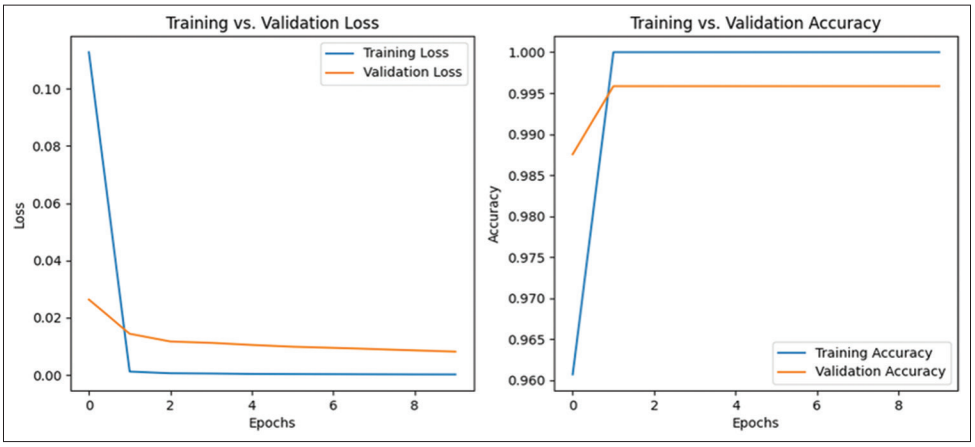

Fig. 10. Accuracy and loss for the DistilBERT model with the TCGA dataset.

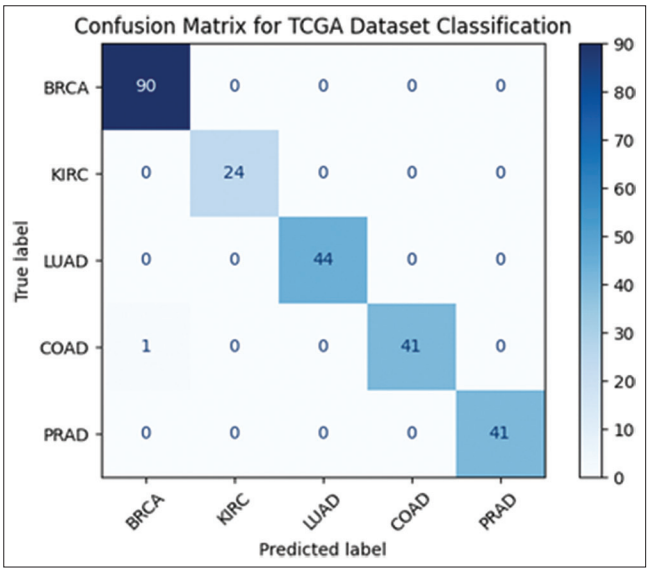

Fig. 11. Confusion matrix of the DistilBERT model with the TCGA dataset.

profile that simplifies classification. When we compared it to previous studies, the proposed model achieved higher accuracy, demonstrating its robustness and effectiveness in cancer classification. The results indicate that integrating transformer-based architectures can enhance predictive accuracy, making it a promising approach for gene expression analysis. By comparison with recent works, Table 9 indicates the comparison between those papers referenced with the proposed method. One limitation of our study model was that it was trained on specific GEO and TCGA datasets, potentially limiting its direct applicability to unseen cancer types or data modalities. The interpretability of the deep learning model also requires further investigation.

7. CONCLUSION

Gene expression profiling for early cancer diagnosis is a new strategy that is intended to help with the early detection and treatment of several types of cancer. In this research, we proposed a DistilBERT model as a multi-class classifier to classify different types of cancer from a variety of sources, including a cancer dataset. We obtained lung and ovarian cancer from GEO, which provides microarray datasets, and used each one separately. We downloaded (BRCA, KIRC, COAD, LUAD, and PRAD) from TCGA, which offered RNA-Seq datasets, which were then merged to create a substantial dataset for cancer classification. We employed a self-attention mechanism to select important features in the dataset and compare the performance of our proposed method with other models and techniques that are used in ML to classify cancer types. We conclude that our proposed model achieved the highest performance compared to other ML methods and techniques. As a result, our proposed approach can accurately categorize all of the observed positive cancer cases. The suggested model can improve early identification of cancer susceptibility, guiding early intervention decisions and ultimately improving survival rates. The suggested model surpasses others across all datasets, achieving the highest classification accuracy: 97.56% for lung cancer, 100% for ovarian cancer, and 99.504% for the TCGA dataset, which includes five types of cancer. In the future, we plan to boost the quality of gene expression data and use metaheuristic optimization alongside deep learning to take our performance to the next level, and we will explore metaheuristic optimization for feature selection and hyperparameter tuning. We also aim to evaluate the model on broader datasets.

REFERENCES

[1] F. Aldi, F. Hadi, N. A. Rahmi and S. Defit. "Standardscaler's potential in enhancing breast cancer accuracy using machine learning". *Journal of Applied Engineering and Technological Science*, vol. 5, no. 1, pp. 401-413, 2023.

[2] L. Rukhsar, W. H. Bangyal, M. S. Ali Khan, A. A. Ag Ibrahim, K. Nisar and D. B. Rawat. "Analyzing RNA-seq gene expression data using deep learning approaches for cancer classification". *Applied Sciences,* vol. 12, no. 4, p. 1850, 2022.

[3] N. Tabassum, M. A. S. Kamal, M. Akhand and K. Yamada. "Cancer classification from gene expression using ensemble learning with an influential feature selection technique". *BioMedInformatics,* vol. 4, no. 2, pp. 1275-1288, 2024.

[4] M. L. R. AbdElNabi, M. Wajeeh Jasim, H. M. El-Bakry, M. Hamed N. Taha and N. E. M. Khalifa. "Breast and colon cancer classification from gene expression profiles using data mining techniques". *Symmetry*, vol. 12, no. 3, p. 408, 2020.

[5] H. AlShamlan and H. AlMazrua. "Enhancing cancer classification through a hybrid bio-inspired evolutionary algorithm for biomarker gene selection". *Computers, Materials and Continua,* vol. 79, no. 1, pp. 675-694, 2024.

[6] W. Ali and F. Saeed. "Hybrid filter and genetic algorithm-based feature selection for improving cancer classification in high-dimensional microarray data". *Processes,* vol. 11, no. 2, p. 562, 2023.

[7] M. Khalsan, L. R. Machado, E. S. Al-Shamery, S. Ajit, K. Anthony, M. Mu and M. O. Agyeman. "A survey of machine learning approaches applied to gene expression analysis for cancer prediction". *IEEE Access,* vol. 10, pp. 27522-27534, 2022.

[8] R. K. Singh and M. Sivabalakrishnan. "Feature selection of gene expression data for cancer classification: A review". *Procedia Computer Science*, vol. 50, pp. 52-57, 2015.

[9] F. Alharbi and A. Vakanski. "Machine learning methods for cancer classification using gene expression data: A review". *Bioengineering*, vol. 10, no. 2, p. 173, 2023.

[10] S. Gupta, M. K. Gupta, M. Shabaz and A. Sharma. "Deep learning techniques for cancer classification using microarray gene expression data". *Frontiers in Physiology*, vol. 13, p. 952709, 2022.

[11] M. Mohammed, H. Mwambi, I. B. Mboya, M. K. Elbashir and B. Omolo. "A stacking ensemble deep learning approach to cancer type classification based on TCGA data". *Scientific Reports*, vol. 11, no. 1, p. 15626, 2021.

[12] D. Mukhopadhyay, D. D. Phanord, R. J. Dalpatadu, L. P. Gewali and A. K. Singh. "ML classification of cancer types using high dimensional gene expression microarray data". Preprints. 2024.

[13] B. Büyüköz, A. Hürriyetoğlu and A. Özgür. "Analyzing ELMo and DistilBERT on Socio-political News Classification". In: *Proceedings of the Workshop on Automated Extraction of Socio-political Events from News 2020*. Marseille, France, pp. 9-18, 2020.

[14] Y. Wu, Z. Jin, C. Shi, P. Liang and T. Zhan. "Research on the application of deep learning-based BERT model in sentiment analysis". *arXiv preprint arXiv:2403.08217,* 2024.

[15] S. Jamshidi, M. Mohammadi, S. Bagheri, H. E. Najafabadi, A. Rezvanian, M. Gheisari, M. Ghaderzadeh, A. S. Shahabi and Z. Wu. "Effective text classification using BERT, MTM LSTM, and DT". *Data and Knowledge Engineering*, vol. 151, p. 102306, 2024.

[16] Y. Ji, Z. Zhou, H. Liu and R. V. Davuluri. "DNABERT: Pre-trained bidirectional encoder representations from transformers model for DNA-language in genome". *Bioinformatics*, vol. 37, no. 15, pp. 2112-2120, 2021.

[17] E. C. Garrido-Merchan, R. Gozalo-Brizuela and S. Gonzalez-Carvajal. "Comparing BERT against traditional machine learning models in text classification". *Journal of Computational and Cognitive Engineering*, vol. 2, no. 4, pp. 352-356, 2023.

[18] V. Dogra, A. Singh, S. Verma, Kavita, N. Jhanjhi and M. Talib. "Analyzing DistilBERT for Sentiment Classification of Banking Financial News". In: *Intelligent Computing and Innovation on Data Science: Proceedings of ICTIDS 2021*. Springer, Singapore, pp. 501-510, 2021.

[19] S. Sucharita, B. Sahu and T. Swarnkar. "Efficient Gene expression data analysis using ES-DBN for microarray cancer data classification". *EAI Endorsed Transactions on Pervasive Health and Technology*, vol. 10, pp. 1-12, 2024.

[20] H. Hijazi and C. Chan. "A classification framework applied to cancer gene expression profiles". *Journal of Healthcare Engineering*, vol. 4, no. 2, pp. 255-283, 2013.

[21] T. Thakur, I. Batra, A. Malik, D. Ghimire, S. H. Kim and A. S. Hosen. "RNN-CNN based cancer prediction model for gene expression". *IEEE Access*, vol. 11, pp. 131024-131044, 2023.

[22] I. Guyon, J. Weston, S. Barnhill and V. Vapnik. "Gene selection for cancer classification using support vector machines". *Machine Learning*, vol. 46, pp. 389-422, 2002.

[23] Y. Wei, M. Gao, J. Xiao, C. Liu, Y. Tian and Y. He. "Research and implementation of cancer gene data classification based on deep learning". *Journal of Software Engineering and Applications*, vol. 16, no. 6, pp. 155-169, 2023.

[24] Y. Li, K. Kang, J. M. Krahn, N. Croutwater, K. Lee, D. M. Umbach and L. Li. "A comprehensive genomic pan-cancer classification using the cancer genome atlas gene expression data". *BMC Genomics*, vol. 18, p. 508, 2017.

[25] P. García-Díaz, I. Sánchez-Berriel, J. A. Martínez-Rojas and A. M. Diez-Pascual. "Unsupervised feature selection algorithm for multiclass cancer classification of gene expression RNA-Seq data". *Genomics*, vol. 112, no. 2, pp. 1916-1925, 2020.

[26] L. P. Chen. "Classification and prediction for multi-cancer data with ultrahigh-dimensional gene expressions". *PLos One*, vol. 17, no. 9, p. e0274440, 2022.

[27] A. Das, N. Neelima, K. Deepa and T. Özer. "Gene selection based cancer classification with adaptive optimization using deep learning architecture". *IEEE Access*, vol. 12, pp. 62234–62255, 2024.

[28] A. Yaqoob, N. K. Verma and R. M. Aziz. "Optimizing gene selection and cancer classification with hybrid sine cosine and cuckoo search algorithm". *Journal of Medical Systems*, vol. 48, no. 1, p. 10, 2024.

[29] S. Tarek, R. Abd Elwahab and M. Shoman. "Gene expression based cancer classification". *Egyptian Informatics Journal*, vol. 18, no. 3, pp. 151-159, 2017.

[30] S. Aburass, O. Dorgha and J. Al Shaqsi. "A hybrid machine learning model for classifying gene mutations in cancer using LSTM, BiLSTM, CNN, GRU, and GloVe". *Systems and Soft Computing*, vol. 6, p. 200110, 2024.

[31] J. A. Martínez Logreira. "*Machine learning-based cancer classification using gene expression data*", (Master's thesis). Universidad de los Andes, Bogotá, Colombia, 2020.

[32] F. Neutatz, B. Chen, Z. Abedjan and E. Wu. "From cleaning before ML to cleaning For ML". *IEEE Data Engineering Bulletin*, vol. 44, no. 1, pp. 24-41, 2021.

[33] L. Huang, J. Qin, Y. Zhou, F. Zhu, L. Liu and L. Shao. "Normalization

techniques in training dnns: Methodology, analysis and application". *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 8, pp. 10173-10196, 2023.

[34] J. Sun and Y. Xia. "Pretreating and normalizing metabolomics data for statistical analysis". *Genes and Diseases*, vol. 11, no. 3, p. 100979, 2024.

[35] R. Dang and W. Yu. "Standard deviation effect of average structure descriptor on grain boundary energy prediction". *Materials*, vol. 16, no. 3, p. 1197, 2023.

[36] Z. Huo, G. Du, F. Luo, Y. Qiao and J. Luo. "D-MSCD: Mean-standard deviation curve descriptor based on deep learning". *IEEE Access*, vol. 8, pp. 204509-204517, 2020.

[37] R. Pramanik, B. Banerjee and R. Sarkar. "MSENet: Mean and standard deviation based ensemble network for cervical cancer detection". *Engineering Applications of Artificial Intelligence*, vol. 123, p. 106336, 2023.

[38] Y. Chen, X. Kou, J. Bai and Y. Tong. "Improving bert with self-supervised attention". *IEEE Access*, vol. 9, pp. 144129-144139, 2021.

[39] B. Cui, Y. Li, M. Chen and Z. Zhang. "Fine-tune BERT with Sparse Self-attention Mechanism". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*). Association for Computational Linguistics, Hong Kong, China, pp. 3548-3553, 2019.

[40] B. Ghojogh and A. Ghodsi, "*Attention mechanism, transformers, BERT, and GPT: Tutorial and survey*," [Preprint] 2020.

[41] Y. Hao, L. Dong, F. Wei and K. Xu. "Self-attention attribution: Interpreting information interactions inside transformer". *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, vol. 35, no. 14, pp. 12963-12971.

[42] J. Shobana and M. Murali. "An improved self attention mechanism based on optimized BERT-BiLSTM model for accurate polarity prediction". *The Computer Journal*, vol. 66, no. 5, pp. 1279-1294, 2023.