e-Prime - Advances in Electrical Engineering, Electronics and Energy 7 (2024) 100449

Contents lists available at ScienceDirect

e-Prime - Advances in Electrical Engineering, Electronics and Energy

journal homepage: www.elsevier.com/locate/prime

Optimal gene therapy network: Enhancing cancer classification through advanced AI-driven gene expression analysis

Tulasi Raju Nethala [a,*], Bidush Kumar Sahoo [a], Pamidi Srinivasulu [b]

[a] Department of Computer Science and Engineering, GIET University, Gunupur Odisha, 765022, India
[b] Department of Computer Science and Engineering, Swarandhra College of Engineering and Technology, Seetharampuram, Narsapur, Andhra Pradesh, 534280, India

ARTICLE INFO

Keywords:
Gene therapy
Gene expression sequences
Light gradient boosting
Harris hawk optimization
Deep learning
Convolution neural network

ABSTRACT

Gene therapy is an advanced medical approach that aims to find solutions for various cancers by identifying optimal gene expressions. In this context, computer-aided detection of gene expressions becomes a research challenge, where artificial intelligence methods are employed to classify cancer types. However, traditional machine learning models must be improved for accurately classifying cancers, leading to unsatisfactory quantitative performance. Therefore, this work implemented the optimal gene therapy network (OGT-Net) for identifying the different types of cancers from the gene expression sequences. Initially, the dataset pre-processing operation normalizes the dataset, which maintains the uniform nature of all records in the dataset. Then, the light gradient boosting model (LGBM) extracts the correlated features from the pre-processed dataset, which contains the relationship among the pre-processed gene expression data. In addition, interrupt-based Harris Hawk optimization (IHHO) extracts the optimal features from LGBM data, decreasing the total number of features by removing redundant gene sequences. Then, a customized deep learning convolution neural network (DLCNN) is used to categorize diseases using gene expression datasets based on lymphography, colon, lung, ovarian, and prostate cancers. The simulation results reveal that the proposed OGT-Net improved performance on various datasets compared to existing approaches, with an average accuracy of 91.128 %, precision of 90.836 %, recall of 91.25 %, and F1-score of 90.7 %.

Introduction and related work

Gene expression analysis has revolutionized our understanding of various cancers, including colon, lung, ovarian, prostate, and lymphography. By examining gene expression patterns, researchers can unravel the underlying molecular mechanisms driving these cancers [1], identify potential biomarkers, and develop targeted therapies. The impact of colon cancer on the health of people all over the world is substantial. Even today, lung cancer is the most common kind of cancer that results in mortality worldwide [2]. Gene analysis has enabled the identification of molecular subtypes within lung cancer, providing valuable information on cancer characteristics, prognosis, and treatment response. Ovarian cancer is a challenging disease to diagnose and treat due to its heterogeneity. One of the malignancies that affects men most often and at the highest rate is prostate cancer. Lymphography is a technique used to assess lymph nodes and lymphatic vessels for potential cancer metastasis [3,4]. Gene expression analysis can enhance the characterization of these findings and provide insights into cancer progression and metastatic potential. It is critical in identifying different subtypes, determining patient prognosis, and guiding treatment selection. In recent years, artificial intelligence (AI) [5], machine learning (ML) [6], and deep learning (DL) [7] models have emerged as powerful tools for analyzing gene expression data and improving cancer diagnosis, prognosis, and treatment strategies. Traditional cancer classification methods rely on subjective histopathological features or genetic mutation interpretations. Gene expression-based classification offers a more objective and accurate approach by simultaneously analyzing the activity levels of thousands of genes. AI, ML, and DL can enhance the accuracy of classification algorithms, leading to improved diagnostic capabilities. Gene expression profiling generates large-scale datasets with thousands of genes and limited samples. Manual analysis of such high-dimensional data is challenging and prone to errors. AI, ML, and DL algorithms excel in handling complex, multidimensional datasets, enabling efficient analysis and classification of gene expression patterns [8].

Gene expression-based cancer classification can uncover novel

* Corresponding author.
E-mail addresses: tulasiraju.nethala@giet.edu (T.R. Nethala), bidushsahoo@giet.edu (B.K. Sahoo), drspamidi@gmail.com (P. Srinivasulu).

https://doi.org/10.1016/j.prime.2024.100449
Received 27 December 2023; Received in revised form 25 January 2024; Accepted 3 February 2024
Available online 14 February 2024

biomarkers associated with specific cancer subtypes or disease progression. Identifying reliable biomarkers can aid in early detection, risk assessment, and monitoring of therapeutic response. AI, ML, and DL methods can expedite the discovery of significant gene expression patterns, leading to the identification of potential biomarkers. Integrating AI, ML, and DL algorithms into clinical practice can provide valuable decision-support tools for oncologists and pathologists [9]. By accurately classifying cancer samples based on gene expression data, these algorithms can assist in treatment planning, patient stratification, and predicting patient outcomes, ultimately improving patient care and outcomes. So, applying AI, ML, and DL techniques to gene expression-based cancer classification offers numerous advantages, including improved diagnostic accuracy, personalized treatment strategies, efficient high-dimensional data analysis, biomarker discovery, and clinical decision support. Addressing this problem can profoundly impact cancer research, patient care, and the development of precision medicine approaches.

Further, building robust and generalizable cancer classification models is essential for their clinical applicability. Models trained on one dataset perform poorly on different datasets due to variations in experimental conditions [10], platforms, or patient populations. It is crucial to develop models that can be generalized across different datasets and cancer types for their widespread adoption and clinical utility. Biological variability, such as genetic mutations, cancer heterogeneity, and differences in gene expression patterns across individuals, can impact the accuracy of cancer classification models. Developing models robust to biological variability and can handle inter-patient and intra-patient heterogeneity is critical for achieving reliable and accurate classifications. The novel contributions of this work are as follows:

- Development of OGT-Net for identification of cancers from colon, lung, ovarian, prostate, and lymphography-based gene expression datasets.
- Dataset normalization ensures that all gene expression records are standardized, enhancing the consistency and comparability of the data.
- The LGBM identifies correlated features in the pre-processed gene expression dataset, capturing meaningful relationships among the genes.
- The IHHO eliminates repeated gene sequences, reducing the number of features and improving computational efficiency.
- The customized DLCNN is specifically designed to classify various cancer disorders from gene expression datasets accurately.

The remaining parts of the paper are structured as follows: Section 2 contains a literature survey on various gene expression-based cancer detection approaches. Section 3 focused on a detailed analysis of OGT-Net with LGBM, IHHO, and DLCNN operations. Section 4 focused on the simulation results of OGT-Net with comparative analysis. Finally, Section 5 concludes the article with future enhancement possibilities.

Review of previous studies

This section contains the related works of various existing gene expression-based cancer classification models. In [11] authors suggested the GECC—Net, which improves the classification performance that utilizes fuzzy ranking in conjunction with a multi-kernel support vector machine (SVM). A subset of genes contributing to cancer categorization was selected using fuzzy ranking. The multi-kernel SVM is employed to classify the samples using the selected gene subset. An ensemble-based method for the categorization of gene expression was developed in [12]. The method used a fuzzy weighted SVM (Fuzzy SVM) and an extreme learning machine (ELM). The fuzzy-weighted SVM-ELM is an extension of the traditional ELM that incorporates fuzzy logic to handle uncertainty in gene expression data. However, these models can require substantial computational properties, and because of this, it is difficult to scale them and deploy them in contexts where resources are limited. In [13] authors provided a modified version of the K-Nearest Neighbors (KNN) approach for gene expression-based cancer categorization. The smallest modified KNN (SMKNN) and largest modified KNN (LMKNN) algorithms incorporate a feature selection mechanism based on mutual information to identify informative genes. However, these models need significant retraining to adapt and learn from new data, hindering their ability to improve continuously.

When attempting to anticipate the temporal dynamics of the genes, Monti et al. [14] relied on an RNN with dual attention. Next, this work concentrated on the RNN's attention mechanism and demonstrated that its graph features enable one to use graph theory tools to identify various GRN architectures hierarchically. However, this model has the potential to perpetuate biases that are already present in the data that it is trained on, which leads to biased results and judgments that are not fair. Zhang et al. [15] developed transformer-based Gene Expression Modeling (T-GEM), a unique interpretable DL architecture. This work showed how the T-GEM model was used to forecast cancer-related symptoms based on gene expression, including categorizing kinds of immune cells and predicting cancer type. However, these models are often considered "black boxes" as they are difficult to interpret and understand, making it challenging to identify how and why they make certain predictions or decisions. Mostavi et al. [16] developed CNN models created using unstructured gene expression inputs. The 1D-CNN, the 2D-Vanilla-CNN, and the 2D-Hybrid-CNN are three separate forms of CNNs that use various convolution algorithms and gene embedding methods. Each of these CNN models is referred to as a CNN model. However, these models heavily rely on high-quality data for training. Limited or biased datasets can result in models that could be more representative and accurate in real-world scenarios.

Haam et al. [17] discovered gene expression profiles associated with the tumor nonimmune microenvironment, and four ML classifiers found that this signature had a high ability to predict brain metastasis. However, these models heavily rely on high-quality data for training. Limited or biased datasets can result in models that could be more representative and accurate in real-world scenarios. Liu et al. [18] presented a reliable early lung cancer diagnosis method. Therefore, creating a deep neural network model is crucial for the early detection of lung cancer. However, the dimensionality curse and uneven data are the fundamental difficulties in mining gene expression databases. However, these models handling large datasets or high-traffic applications can be a significant challenge due to computational limitations and the need for efficient algorithms. Soto et al. [19] examined the most popular machine and DL classification algorithms for the cancers included in the T11 tumor database. Using k-fold cross-validation, this study identified tumors with accuracy ranging from 90.60 % (logistic regression) to 94.4 % for CNNs. However, these were vulnerable to adversarial attacks or exploitation, posing data security and privacy risks. Ke et al. [20] focused on selecting filter-wrapper genes via swarm optimization. In the first step, only a select few genes that rank in the top n are considered, known as the filter step, which results in decreased data. However, the computational complexity of these models is high.

Hamzeh et al. [21] developed a technique to find gene sets that distinguish between distinct laterality classes. The resulting genes have been discovered to correlate with disease development substantially. However, these models can be sensitive to variations or perturbations in input data, leading to unexpected and potentially incorrect outputs. Ensuring robustness in noisy or incomplete data is a significant challenge. Using ML methods, including SVM and random forest, Yuan et al. [22] analyzed the gene expression patterns of the lung and lung samples acquired from Gene Expression omnibus. After that, an effective method of feature selection known as monte carlo feature selection was used to evaluate the profiles. However, this method requires higher computational complexity. Díaz et al. [23] used the ELM algorithm and analysis accuracy, discovering that balanced databases had higher accuracy than unbalanced ones. A weighted majority vote system among the feature

groups was used to determine the final categorization. However, this method requires lower performance. Chen et al. [24] suggested employing overlapping conventional feature selection strategies for cancer classification and biomarker development. Using the overlapping technique, A random forest was used to confirm the genes chosen. However, this method requires higher computational complexity. Divate et al. [25] examined the performance of a trustworthy neural network model for reliably identifying cancers using gene expression data as the input. In addition, they generated a list of gene signatures that could be beneficial for creating biomarker panels, which are used to pinpoint the tissue in which cancer first began. However, this method resulted in lower accuracy with higher computational complexity.

In [26], the authors introduced an innovative gene selection approach termed the Sine Cosine and Cuckoo Search Algorithm (SCACSA). This hybrid method was designed to complement well-known ML classifiers, such as SVM. The authors meticulously assessed the hybrid gene selection algorithm's performance using a breast cancer dataset, comparing its effectiveness to other feature selection methods. Moving on to [27], the authors emphasized the superior structural representation of biological systems through graphs in biomedical data. While acknowledging the existence of graph neural network (GNN) based multi-omics integrative methods, they highlighted three common disadvantages. Firstly, many of these methods exclusively employed either inter-omics or intra-omic connections. Secondly, they often focused on a single GNN layer, such as a graph convolution network (GCN) or graph attention network (GAT). Thirdly, these methods were seldom tested on complex classification tasks like cancer molecular subtypes. In [28], a novel end-to-end DL approach named DeepGene Transformer was proposed. This approach addressed the complexity of high-dimensional gene expression data through a multi-head self-attention module. The method aimed to identify relevant biomarkers across multiple cancer subtypes without requiring feature selection as a prerequisite for the current classification algorithms. Moving on to [29], the authors presented a classification method to understand the convergence of training deep neural networks (DNN). They assumed that inputs did not degenerate, the network was over-parameterized, and the number of hidden neurons was sufficiently large. The authors utilized DNN for classifying gene expression data, specifically focusing on a dataset containing bone marrow expressions from 72 leukemia patients. In [30], the authors addressed a neglected area in swarm-optimization-based filter-wrapper gene selection. They proposed a method called Population Initialization based on Ranking Criteria (PIRC) to transform the population initialization of genetic algorithm (GA) and ant colony optimization (ACO), referred to as PIRCGA and PIRCACO, respectively. The experiment was conducted on 17 microarray expression datasets, comparing two groups: IG-GA vs. IG-PIRCGA and IG-ACO vs. IG-PIRCACO.

Research gaps

The major research gap is the need for robustness and generalizability of models. Many existing models are often trained and evaluated on specific datasets or cancer types, limiting their ability to generalize to diverse and unseen data. Ensuring that ML and DL models can generalize across different cancer types, patient populations, and datasets is crucial for their broader applicability in clinical settings. Computational efficiency and scalability represent another significant research gap. While powerful DL models like transformers have shown promise, their computational demands can be prohibitive, especially in resource-constrained environments such as healthcare facilities. Developing ML and DL models that balance accuracy with computational efficiency is essential for practical deployment in real-world scenarios. Furthermore, addressing biases in the training data and the models is a critical research gap. Biases in gene expression data or how models are trained can lead to skewed predictions and unfair outcomes, potentially impacting vulnerable populations. Research efforts should focus on developing methods to detect and mitigate biases in both data and models to ensure fairness and equity in cancer detection applications. The dynamic nature of gene expression data also presents a research gap. Many existing models need help to adapt to changes in gene expression patterns over time or in response to different treatments. Enhancing the adaptability of ML and DL models to evolving data is essential for their utility in dynamic clinical environments. Lastly, collaborative, and reproducible research practices are essential for advancing the field. Establishing standardized benchmarks, datasets, and evaluation metrics can facilitate fair comparisons between different models and promote reproducibility in gene expression-based cancer detection research. Bridging these research gaps will contribute to developing more reliable, scalable, and interpretable ML and DL models for gene expression-based cancer detection.

Novelty of work

The research makes significant strides in overcoming existing research gaps by introducing novel contributions that address key challenges in gene expression-based cancer detection. Firstly, the implementation of the LGBM represents a crucial advancement. The LGBM can identify correlated features within the pre-processed gene expression dataset, thereby capturing meaningful gene relationships. This is essential for enhancing the interpretability of the model, a critical research gap, as it provides insights into the intricate interactions between genes that contribute to cancer categorization. The ability of the LGBM to discern and utilize these correlations contributes to a more transparent and understandable machine-learning approach. Secondly, the IHHO mechanism introduced in the research addresses computational efficiency concerns by eliminating repeated gene sequences. This innovative approach reduces the number of features in the dataset, consequently enhancing computational efficiency. It is a vital contribution to overcoming the challenge of scalability and resource constraints, making the proposed model more suitable for real-world applications, particularly in healthcare settings with limited computational resources. Lastly, the research introduces a customized DLCNN designed to classify various cancer disorders from gene expression datasets accurately. This tailored approach acknowledges the diversity of cancer types and aims to develop a model that can effectively categorize them. Customizing the DLCNN is a crucial step toward achieving robustness and generalizability, addressing the research gap related to the limited applicability of existing models to diverse cancer types. By focusing on the nuanced characteristics of different cancers, the DLCNN contributes to more accurate and reliable predictions in gene expression-based cancer classification.

Proposed methodology

AI, ML, and DL models are often perceived as black boxes, making understanding, and interpreting their decisions challenging. In the context of cancer classification, it is crucial to have models that provide interpretability and explainability, allowing clinicians and researchers to understand the underlying biological mechanisms and justify the classification results. Developing interpretable models while maintaining high accuracy is an ongoing research challenge. The real-time workflow, depicted in Fig. 1, encompasses the journey from patient sample collection to the outcome communicated to the patient. The process begins with acquiring gene expression datasets stored on the Real-Time Hospital Server. These datasets have been meticulously verified and approved by domain experts, ensuring the genetic information's reliability and accuracy. The data serves as the foundation for training the proposed OGT-Net, a deep-learning model designed to isolate and characterize human gene DNA. The training of OGT-Net involves feeding it with the gene expression datasets, allowing the model to learn intricate patterns and correlations within the genetic data. The trained model is then saved and deployed on the real-time

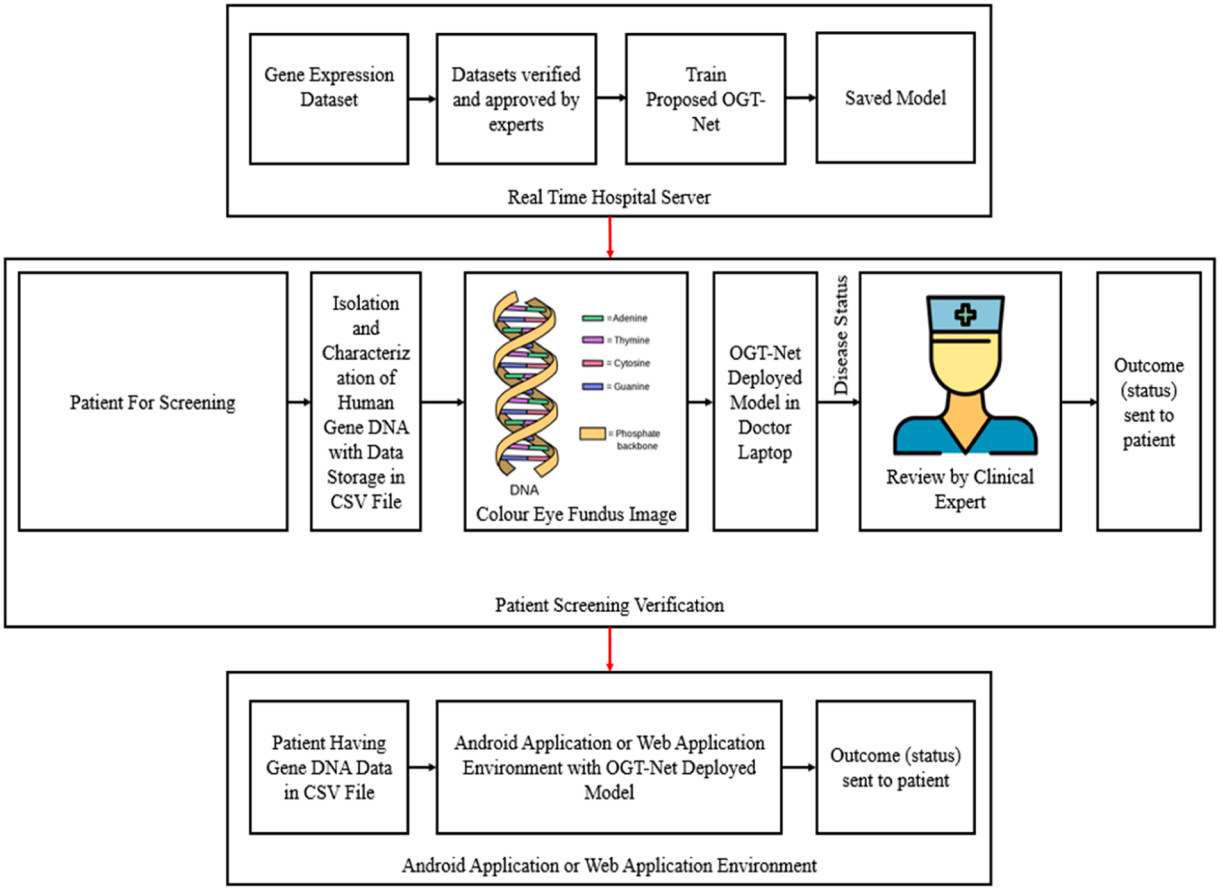

Fig. 1. Application Scenario of OGT-Net.

hospital server for subsequent use in patient screening.

Patient screening and verification represent a crucial stage in the healthcare pipeline. It involves the collection of genetic samples from patients, a process akin to isolating and characterizing human gene DNA. These samples are meticulously processed, and the resulting genetic information is stored in a structured CSV file for further analysis. This CSV file becomes a central component in the application scenario as input to the OGT-Net deployed on the doctor's laptop. The OGT-Net, armed with the knowledge gained during training, rigorously analyzes the genetic data, providing insights into the patient's disease status. The deployment of the OGT-Net on the doctor's laptop facilitates real-time decision-making, allowing for prompt and accurate patient assessments. The outcome of the analysis, reflecting the disease status, undergoes a critical review by a clinical expert. This step ensures that the conclusions drawn by OGT-Net align with healthcare professionals' nuanced understanding and expertise. The collaborative effort between advanced computational models and human expertise enhances diagnostic accuracy. Once the clinical expert reviews the results, the outcome, representing the patient's disease status, is communicated back to the patient. This transparent and communicative approach ensures that patients are informed about their health status, fostering a sense of involvement, and understanding in their healthcare journey. The integration continues within the confines of the hospital server.

The patient's gene DNA data and the OGT-Net deployed model extend to an Android or Web Application Environment. This environment provides an accessible interface for both patients and healthcare providers. Patients can easily upload their gene DNA data in CSV format through the application. Within this application environment, the OGT-Net deployed model processes the uploaded genetic data in real time, generating outcomes that are promptly sent back to the patients. Utilizing an android or web application environment streamlines the entire process, making it user-friendly and accessible to a broader demographic. In essence, this integrated system signifies a transformative approach to patient care. The real-time nature of the system, coupled with the transparent communication of outcomes to patients, empowers individuals to participate in their healthcare journey actively.

Fig. 2 shows the proposed OGT-Net, a prominent research solution to gene sequence analysis. Initially, normalizing the gene expression dataset ensures that all records have a uniform nature, eliminating potential biases caused by variations in gene expression levels. This step improves the reliability and accuracy of subsequent analyses and classification models. Then, using LGBM to extract correlated features from the pre-processed dataset is a significant contribution. The LGBM is a powerful machine-learning algorithm that efficiently handles high-dimensional data. By identifying and leveraging the correlations between gene expression features, LGBM helps capture the underlying relationships within the data, leading to more informative and discriminative features for cancer classification. The basic HHO suffers from over-fitting and higher execution time due to default 1000 iterations and 50 populations. So, to reduce these complexities, basic HHO is interrupted automatically and formed as IHHO. The IHHO stops iterations when it reaches maximum accuracy. The IHHO used to extract optimal features from the LGBM data is a novel contribution. The hunting strategy of Harris hawks served as the basis for developing the IHHO optimization algorithm. By employing interrupt-based strategies, IHHO identifies and eliminates repeated gene sequences, reducing the number of features. This feature selection process improves computational efficiency, reduces overfitting, and enhances the interpretability of the classification models. Finally, a customized DLCNN is used to classify various cancer disorders. Customizing the DLCNN implies that it is specifically designed for cancer classification, considering the unique characteristics of the gene expression datasets associated with colon,

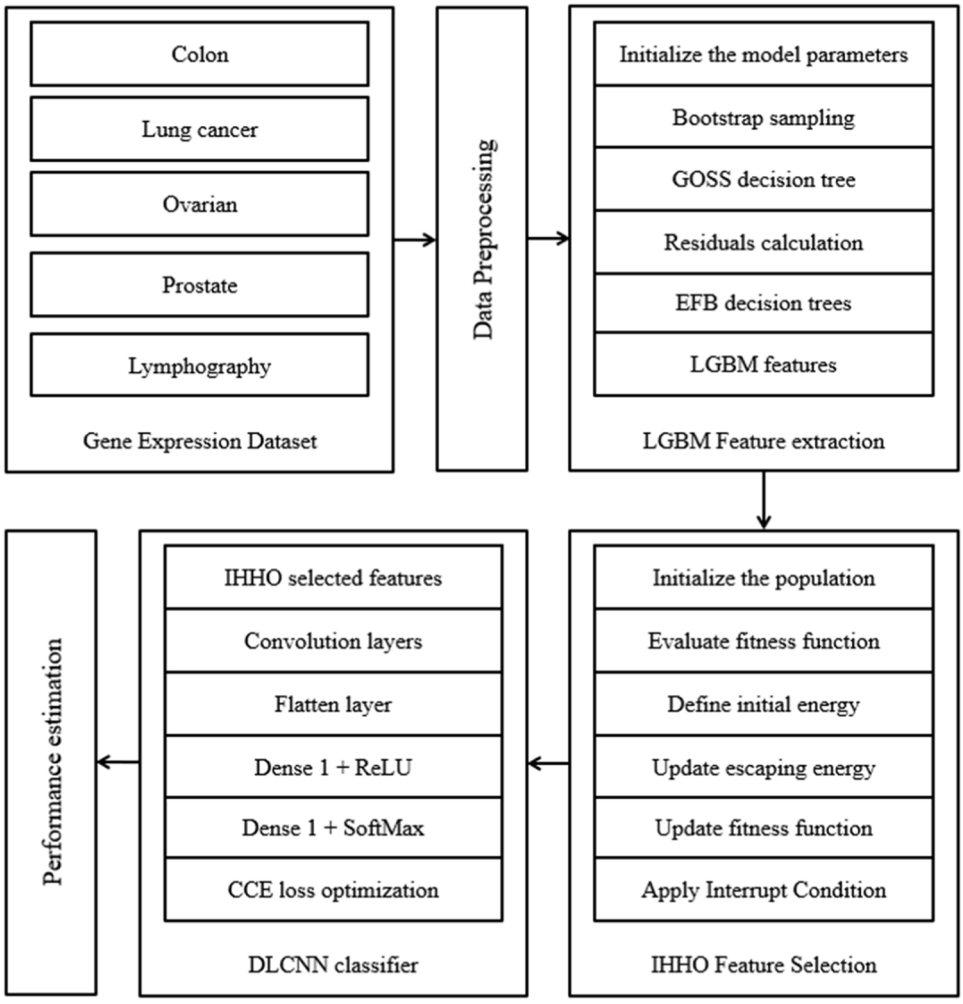

Fig. 2. Proposed OGT-Net block diagram.

lung, ovarian, prostate, and lymphography disorders. This customized approach allows for more accurate and disease-specific classification results, facilitating improved diagnosis and personalized treatment strategies.

Preprocessing

The preprocessing of the data includes the normalization step, which involves adjusting the gene values such that there are less significant differences between the various genes. The min–max approach is used in this study for normalization. This method re-scales the range of values for each gene to the interval [0,1]. The standard gene value is determined using a set C of instances characterized by a set G of genes $v_{c_i,\ g_i}$, $c_i \in C,$ and $g_j \in G$ are some gene values. Further, the normalized dataset is $\widehat{v}_{c_i,\ g_i}$, which is derived as follows:

$$\widehat{v}_{c_i,g_i} = \frac{v_{c_i,g_i} - \min c_k \in C\left(v_{c_k,g_j}\right)}{\max\left(c_k \in C\left(v_{c_k,g_j}\right)\right) - min\left(c_k \in C\left(v_{c_k,g_j}\right)\right)} \tag{1}$$

Here, the lowest and maximum values of gene $g_j \in G$ across all of the instances $c_k \in C$ are denoted by the notations $min\ (c_k \in C(v_{c_k,\ g_j}))$ and $\max(c_k \in C(v_{c_k,g_j}))$, respectively. The problem of missing values, which is extremely persistent in virtually all experimental datasets, is addressed and solved by this body of work once it has been normalized. To complete the missing values for each gene, this method computes the mean of the existing values for each gene and uses it to fill in the blanks. The set of instances with values for the gene g_j is denoted by the superscript $C^{'}g_j \subset C$, while the set of cases that do not have a value for that gene is denoted by the superscript $C^{'}g_j \subset C$. Then, the value for gene g_j is assigned by this work to each case $C^{'}g_j \subset C$ that does not already have a value for gene g_j. It is done for each gene $g_j \in G$. The next step is to generate the pre-processed dataset, denoted by the acronym $P_{c_i,\ g_i}$.

$$P_{c_i,g_i} = \frac{\sum_{cj \in C_{g_j}} \widehat{v}_{c_i,g_j}}{\left|C_{g_j}\right|} \tag{2}$$

Here, $|C_{g_j}|$ denotes the cardinality of set C_{g_j}.

LGBM feature extraction

Gene expression data in cancer classification often includes a relatively small number of samples and many genes (features). The large dimensionality of this data presents difficulties, particularly in feature selection and dimensionality reduction, and avoiding overfitting. Effective AI, ML, and DL techniques must be applied to handle this complexity and extract meaningful information from the data. The main difference between LGBM and other gradient boosting frameworks is that LGBM expands in a vertical direction means it grows leaf-wise. At the same time, the other algorithms expand horizontally in a level-wise direction. LGBM selects the leaf that produces the least error and maximum efficiency in feature extraction. Table 1 shows the LGBM feature extraction algorithm procedure. Fig. 3 shows the proposed

Table 1
LGBM feature extraction algorithm.

Input: Preprocessed dataset
Output: LGBM features
Step 1: Set the initial values for the model's parameters, which include the number of boosting iterations, the learning rate, and other tree-specific values.
Step 2: Split the training data into multiple subsets, usually using random sampling with replacement (bootstrapping), creating different subsets called "bootstrap samples."
Step 3: Build an initial GOSS decision tree for each bootstrap sample, often called a "weak learner" or "base learner." This tree predicts the target variable based on the selected features.
Step 4: Calculate the residuals (the deviations between the expected and observed values) of the current model's predictions for each data point in the training set.
Step 5: Fit a new GOSS decision tree to the residuals, aiming to capture the patterns and relationships the previous tree failed to capture.
Step 6: Update the model's predictions by adding the predictions from the new decision tree to the previous predictions, weighted by a learning rate. The pace of learning determines how much each tree contributes to the ensemble.
Step 7: Repeat steps 4 to 6 for several boosting iterations, creating an ensemble of EFB decision trees.
Step 8: Obtain the final predicted LGBM features by aggregating the predictions of all EFB trees in the ensemble.

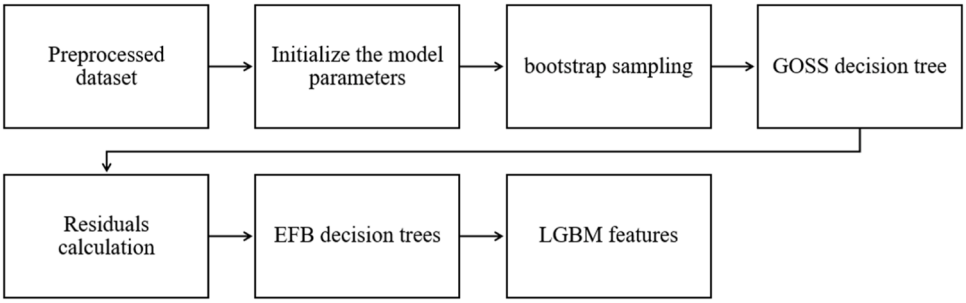

Fig. 3. LGBM feature extraction block diagram.

LGBM feature extraction block diagram. The decision trees that comprise the LGBM serve as the basis for the gradient-boosting architecture. Its objective is to lessen the model's reliance on the user's memory while simultaneously improving the model's efficacy. The LGBM uses two innovative approaches: gradient-based one-side sampling (GOSS) and exclusive feature bundling (EFB).

The LGBM can circumvent the limitations of traditional histogram-based algorithms, such as the gradient-boosting decision tree, with the assistance of these approaches. When computing the amount of information gained, various data instances each play a unique role in the computation. Instances with larger gradients, also known as instances, that have undergone inadequate training, will contribute more to the total knowledge gained. It is because larger gradients indicate that the instances have not been adequately trained. GOSS only drops arbitrarily those instances with tiny gradients to maintain the accuracy of the information gain estimate, and it preserves those instances with big gradients (for example, those greater than a predetermined threshold or those among the top percentiles). It is done to maintain the accuracy of the information gain estimate. When the value of the information gain is spread out across a large range, this technique provides a more accurate estimate of the gain than uniformly random sampling, even when the intended sample rate is kept the same. This is particularly true when the range of possible information gains is broad. This work can create a nearly lossless way to reduce the number of features since high-dimensional data are often sparse.

Further, a sparse feature space has several features that cannot simultaneously assume values other than zero. It indicates that these characteristics cannot coexist with one another. The EFB allows the unique features to be consolidated into a single, risk-free feature at that point. Consequently, the difficulty of creating a histogram change from $O(\#data \times \#feature)$ to $O(\#data \times \#bundle)$, with $\#bundle$ being more complicated than #feature. As a direct consequence, the training framework's speed has accelerated while preserving the previous degree of accuracy. If we have a training set of n occurrences, referred to as $\{x_1, \ldots, x_n\}$, each x_i is a vector with dimension s in space X_s for a training set with n instances. The negative gradients of the loss function concerning the output of the model are denoted by the notation "$\{g_1, \ldots, g_n\}$" throughout each iteration of the LGBM algorithm. "For a training set with n instances," "for a training set with n instances." Within the framework of this GOSS method, the training examples are arranged in descending order based on the absolute values of the gradients that they have. The top-$a > 100$ % instances with the largest gradients are then retained, and LGBM is provided with an instance subset denoted by the letter A. Then, the remaining set A_c consisted of examples with lesser gradients and had a ratio of $(1 - a) \times 100\%$ greater than 100 percent. The LGBM then takes a sample at random from a subset B whose size is greater than $b \times |A_c|$.

Finally, LGBM divided the cases into two groups based on the projected adjustment gain at vector $V_J\ (d)$ across the A versus B subset.

$$\widetilde{V}_j(d) = \frac{1}{n}\left(\frac{\left(\sum_{x_i \in A_l} g_i + \frac{1-a}{b}\sum_{x_i \in B_l} g_i\right)^2}{n_l^j(d)} + \frac{\left(\sum_{x_i \in A_r} g_i + \frac{1-a}{b}\sum_{x_i \in B_r} g_i\right)^2}{n_l^j(d)}\right) \tag{3}$$

Here, $A_l = \{x_i ? A : x_{ij} ? d\}$, $A_r = \{x_i ? A : x_{ij} > d\}$, $B_l = \{x_i ? B : x_{ij} ? d\}$, $B_r = \{x_i ? B : x_{ij} > d\}$, and It is necessary to employ the constant $\frac{1-a}{b}$ to get the total of the gradients across B back to the original size of A_c.

IHHO feature selection

Not all gene data were relevant or informative for cancer classification. Feature selection techniques must be employed to identify the subset of genes with the most discriminatory power. It helps minimize the computational complexity and enhances how easily categorized models can be interpreted. Optimization algorithms can lend a hand in the process of automatically selecting or extracting features based on the findings from the gene expression. The block diagram of the IHHO algorithm is shown in Fig. 4. Further, Table 2 presents a detailed pseudo code of the IHHO algorithm. It is well-known that hawks pursue their prey by tracking, surrounding, and ultimately hitting and killing the animal. The hunting behaviors of hawks provide the foundation for the three separate phases that the mathematical model represents. These stages include exploration, transition between exploration and

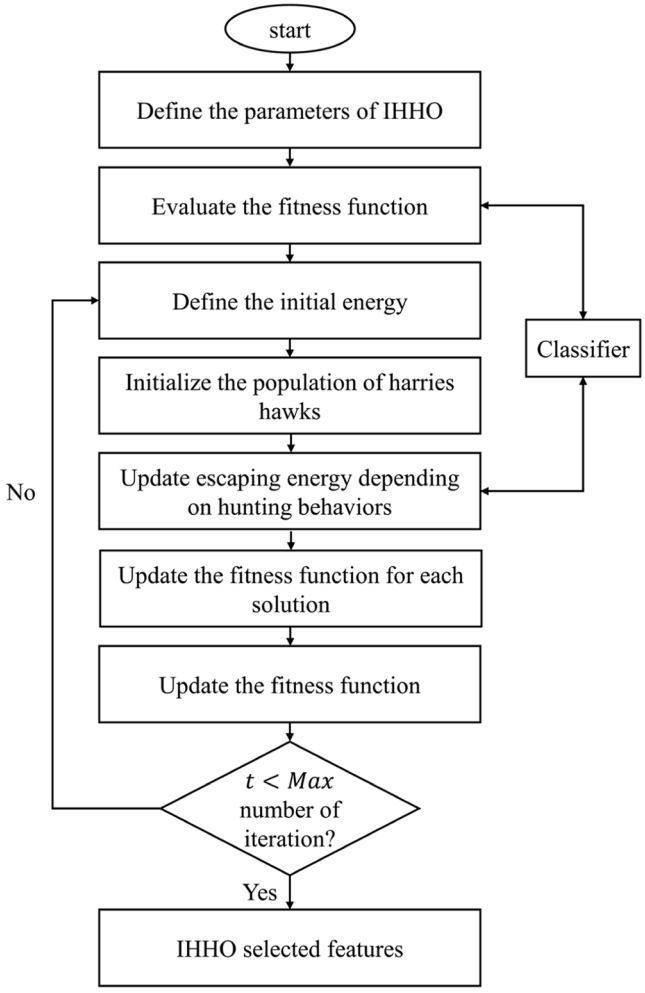

Fig. 4. Block diagram of IHHO.

exploitation, and exploitation. At each stage of the hunt, the Harris's hawks are the candidate solutions, and the prey sought is the best candidate solution (almost the ideal). Harris's hawks use two distinct exploration methods throughout their prey hunt. Candidates for solutions are developed to be as like the prey as is practically feasible, but the solution that most closely matches the prey is the optimal choice. The first step in the location selection process for Harris's hawks is to consider where other species of hawks and the prey they hunt are found in the vicinity. As a part of the second plan, the hawks would wait on various trees that are rather tall. The two approaches were simulated using Eq. (4) with the same probability of obtaining q.

$$x(t+1) = \begin{cases} x_{random}(t) - r_1 |x_{random}(t) - 2r_2(t)| & q \geq 0.5 \\ x_{rabbit}(t) - x_{mean}(t) - r_3(LB + r_4(UB - LB)) & q < 0.5 \end{cases} \quad (4)$$

The location of the hawk during the current iteration is characterized by the vector $x(t)$. In contrast, the hawk's position during the following iteration is signified by the vector $x(t+1)$. The hawk with the identifier $x_{random}(t)$ is randomly picked out of the population. The $x_{rabbit}(t)$ position represents the rabbit position. The integers q, r_1, r_2, r_3, and r_4 were randomly produced within (0,1). The upper and lower limits of the variables are denoted by the notations UB and LB, respectively. Eq. (5) is the representation of the equation that is used in the process of determining the typical distribution of the current population of hawks. The $x_{mean}(t)$ value represents this average position.

$$x_{mean}(t) = \frac{1}{N} \sum_{i=1}^{N} x_i(t) \quad (5)$$

In this situation, the total number of iterations is denoted by t. The location of each hawk during the iteration t is denoted by the x_i (t) variable. The symbol N is used to denote the overall number of hawks. As illustrated in Eq. (6), the algorithm changes its focus from examination to exploitation based on how much energy the rabbit uses when running or attempting to escape.

$$E = 2E_0 \left(\frac{t}{Max\_iter} \right) \quad (6)$$

In this case, E stands for the energy that the prey uses to flee. The starting condition of the energy is denoted by the value E_0, which undergoes unpredictable shifts anywhere within the range (1, 1) after each cycle. Here, E is less than one, and hawks search for more territories to

Table 2
IHHO feature selection pseudo code.

Input: LGBM features
Output: Optimal IHHO features
Step 1: LGBM extracted gene features as input
Step 2: Define the parameters of IHHO
- Max_Iterations: maximum number of iterations
- Population_Size: size of the population=50
- Escape_Rate: rate of energy loss during escaping
- Capture_Rate: rate of energy gain during capturing
- Mutation_Rate: rate of mutation in the population

Step 3: Evaluate the fitness function
- Fitness function evaluates the performance of a solution based on LGBM-extracted gene features

Step 4: Define the initial energy
- Initialize the energy level for each solution in the population

Step 5: Initialize the population of Harris Hawks
- Randomly generate a population of solutions, each representing a set of features

Step 6: Update escaping energy depending on hunting behaviors
- For each solution in the population:
  - Calculate the escaping energy based on hunting behaviors
  - Update the energy level for the solution

Step 7: Update the fitness function for each solution
- Evaluate the fitness of each solution based on the updated set of features

Step 8: Update the fitness function
- Select the top-performing solutions for further optimization

Step 9: t <Max number of iterations? (if Yes, generate output, else No, go back to Step 4)
Step 10: IHHO selected features
- Return the selected features based on the optimization process.

look for rabbits; alternatively, the exploitation stage gets underway. The method determines whether the rabbit was successful in escaping with a probability of $p < 0.5$ or if it was unsuccessful with a probability of $p < 0.5$. The Hawks will also carry out either a gentle or hard siege, with the level of difficulty being determined by the amount of energy the rabbit possesses.

The definition of the soft siege was found in Eqs. (7)–(9).

$$x(t+1) = \Delta x(t) - E|J \cdot x_{rabbit}(t) - x(t)| \tag{7}$$

$$\Delta x(t) = x_{rabbit}(t) - x(t) \tag{8}$$

$$J = 2(1 - random) \tag{9}$$

The value $x(t)$ denotes the difference between the hawk and the rabbit locations at any time. A random value, denoted by J, is utilized in this context to calculate the rabbit's unpredictable leap force. Conversely, the effectiveness of a severe siege was determined using the formula shown in Eq. (10).

$$x(t+1) = x(t) - E|\Delta x(t)| \tag{10}$$

When E is less than 0.5, and p is less than 0.5, a soft siege consisting of multiple quick dives is attempted since the rabbit can evade capturing. The hawks are given the chance to choose the most advantageous dive. Lévy flying is used to mimic the jumping motion of the prey. The following move by the Hawks is analyzed using the formula presented in Eq. (11), which helps assess whether the dive was effective.

$$k = x_{rabbit}(t) - E|J \cdot x_{rabbit}(t) - x(t)| \tag{11}$$

If the preceding dive is unsuccessful, the hawks will dive again according to Eq. (12), which represents the Lévy flight L pattern.

$$z = k + RandomVector \cdot L(dim) \tag{12}$$

The dimension of the problem, often denoted as "dim," is equivalent to the size of the random vector utilized. Eq. (13) was crucial in updating the soft-siege fast dives to their current state.

$$x(t+1) = \begin{cases} k \text{ if } f(k) < f(x(t)) \\ z \text{ if } f(z) < f((x(t))) \end{cases} \tag{13}$$

To determine optimum features with k and z, respectively, Eqs. (12) and (13) are employed in the calculation process. It is a severe siege with successive quick dives after $|E|0.5$ and p0.5 are not adequate for the rabbit to leave since it no extensive has sufficient energy. This situation causes the rabbit to be unable to escape. Eq. (9) is used to get the value of the rabbit's z, and Eq. (14) is used to determine the rabbit's k.

$$k = x_{rabbit}(t) - E|J \cdot x_{rabbit}(t) - x_{mean}(t)| \tag{14}$$

DLCNN classification

It has been shown that DL approaches, particularly CNN, offer significant promise in extracting complex patterns and representations from gene expression data. Compared to a Multi-Layer Perceptron (MLP), a DLCNN includes distinctive convolutional and pooling layers. DLCNN exhibits excellent cost performance, especially with larger datasets, as it benefits from increased model size and performance. Fig. 5 shows the DLCNN model, which contains the convolutional layer, pooling layer, nonlinear activation function such as rectified linear unit (ReLU), and fully connected layer. Typically, data undergoes preprocessing before entering the network through the input layer. The convolutional layer possesses both local and global receptive field characteristics. It efficiently detects correlations between data sample features in the spatial dimensions while maintaining the input shape. Parameter sharing and sparse connections also enable the convolutional layer to calculate the same convolution kernel at different locations, preventing an excessively large parameter size. The pooling layer reduces the convolutional layer's sensitivity to precise location,

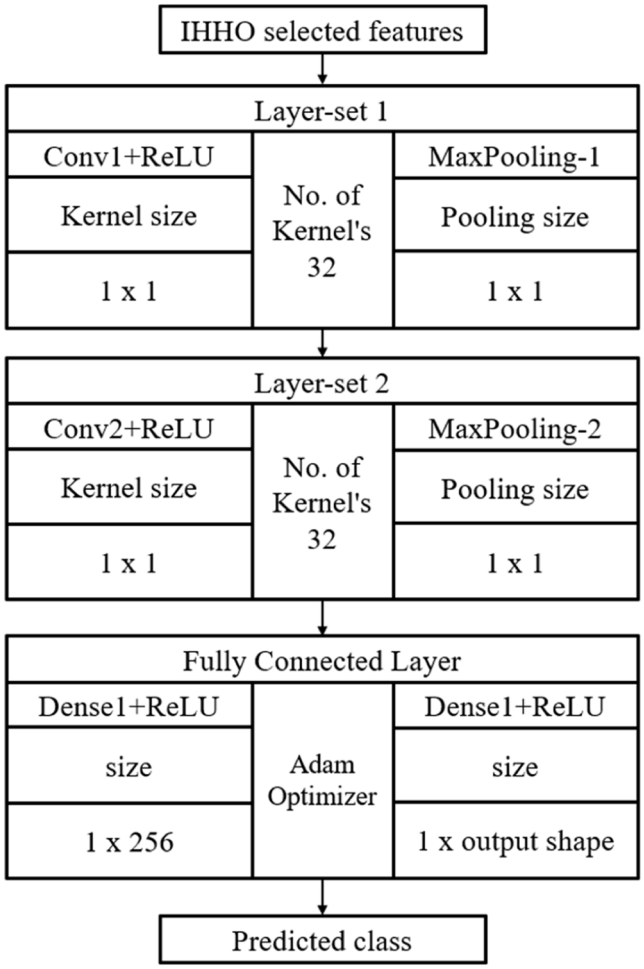

Fig. 5. Proposed DLCNN architecture.

decreasing computational costs. It achieves this by reducing the number of connections between convolutional layers. To a certain extent, DLCNN assures that the input data samples remain invariant regarding their displacement, scaling, and distortion. Subsequently, convolutional and pooling layers process the data in an alternating order. Finally, the fully connected layer categorizes the data and predicts the disease class using the SoftMax classifier.

Results

This section offers a complete analysis of the simulation results utilizing a broad range of performance indicators. In addition, the performance of the proposed OGT-Net is associated with the performance of current techniques utilizing the same dataset in a comparable environment.

Dataset

The proposed OGT-Net is applied to five distinct cancer-related datasets, each representing a different type of cancer—colon, lung, ovarian, prostate, and lymphography. These datasets serve as valuable information repositories, contributing to understanding various cancer types through the lens of gene expression data. Each dataset is characterized by specific features and samples, shedding light on the intricacies of genetic information associated with different cancers. Table 3 provides a detailed analysis of the dataset description with several features and samples. The Colon cancer dataset provides a glimpse into the molecular landscape of colorectal cancer, which contains 2000 features

Table 3
Dataset description.

Dataset	Number of features	Number of samples
Lymphography	18	148
Ovarian cancer	46	253
Colon	2000	62
Lung cancer	4	64
Prostate cancer	8	354

and 62 samples. Moving to the lung cancer dataset comprises four distinct features and 64 samples. The Ovarian cancer dataset, with 46 features and 253 samples, delves into the genetic intricacies of ovarian cancer. The Prostate cancer dataset, characterized by eight features and 354 samples, focuses on the genetic underpinnings of prostate cancer. Lastly, the Lymphography dataset, comprising 18 features and 148 samples, sheds light on the genetic aspects of lymphographic cancer. These datasets provide a more comprehensive view, encompassing a broader array of genetic attributes pertinent to understanding cancer at the molecular level. The increased number of features facilitates a more detailed analysis, contributing to a nuanced comprehension of the genetic landscape associated with cancer cases.

Ablation study

Tables 4 and 5 provide the results of an ablation study performed on the OGT-Net model. The metrics of accuracy, precision, recall, and F1-score were the primary areas of attention. This ablation study intends to examine the influence of several components (LGBM and IHHO) on the performance of OGT-Net across various datasets to provide a recommendation for further research. Here, OGT-Net without LGBM, IHHO variation represents the baseline model without either LGBM or IHHO components. It serves as a reference for evaluating the contributions of these components. The OGT-Net model outperforms this variation by a significant margin in performance measures. The OGT-Net without LGBM feature extraction variation excludes the LGBM component from the OGT-Net model. By comparing it with the complete OGT-Net, the impact of LGBM was analyzed. The results indicate that incorporating LGBM leads to notable improvements in performance across all datasets. The OGT-Net without IHHO feature selection variation removes the IHHO component from the OGT-Net model. It allows us to evaluate the influence of IHHO on the model's performance. The findings reveal that IHHO is crucial in enhancing performance for the OGT-Net model. Finally, the complete OGT-Net model, incorporating both LGBM and IHHO components, serves as a benchmark to measure the overall performance. The OGT-Net achieves the highest performance across all datasets, surpassing all other variations.

Performance comparison

Table 6 presents the results of an analysis of the overall performance of the OGT-Net model, which compares the model's results across many datasets. Accuracy, Precision, Recall, F1-Score, and Execution Time are some of the metrics included in the table, and they provide insights into the efficacy and efficiency of the model. The OGT-Net model achieves high accuracy percentages ranging from 86.66 % to 95.01 % across different datasets, indicating its ability to classify instances in diverse domains correctly. Precision percentages range from 86.11 % to 96.42 %, demonstrating the OGT-Net's capability to make true positive calculations while minimizing false positives. The model successfully recognizes positive cases and avoids false negatives, as shown by the recall percentages, which vary from 90.04 percent to 94.4 percent. The F1 scores vary anywhere from 86.11 % to 94.30 %, which indicates that the OGT-Net model maintains a healthy equilibrium between accuracy and recall metrics.

Table 7 compares the performance of various gene expression analysis methods on the colon dataset. The proposed OGT-Net method exhibits substantial improvements compared to Fuzzy SVM [12], with an accuracy increase of +139.84 %, precision improvement of +368.42 %, recall improvement of +88.80 %, F1-score improvement of +229.30%, and a longer execution time. Compared to SMKNN [13], the proposed OGT-Net method demonstrates significant advancements, achieving an accuracy improvement of +71.46 %, precision enhancement of +208.07 %, recall increase of +115.43 %, F1-score improvement of +161.43 %, and a slightly longer execution time. The proposed OGT-Net method outperforms LMKNN [13], showcasing a remarkable accuracy improvement of +50.77 %, a substantial precision boost of +23.38 %, recall enhancement of +46.12 %, F1-score improvement of +58.99 %, and a slightly longer execution time. Compared to GECC—Net [11], the proposed OGT-Net method achieves notable progress, attaining a significant accuracy improvement of +20.38 %, precision enhancement of +11.11 %, recall increase of +16.15 %, F1-score improvement of +18.58 %, and a slightly longer execution time.

Table 8 compares the performance of various gene expression analysis methods on the lung cancer dataset. The proposed OGT-Net method demonstrates a notable improvement compared to Fuzzy SVM [12], with an accuracy increase of +120.00 %, precision improvement of +340.00 %, recall improvement of +85.70 %, F1-score improvement of +211.33 %, and a longer execution time. Compared to SMKNN [13], the proposed OGT-Net method exhibits a slight improvement, with an accuracy increase of +1.85 %, precision improvement of +2.96 %, recall improvement of +4.68 %, F1-score improvement of +2.47 %, and a longer execution time. The proposed OGT-Net method showcases a slight improvement compared to LMKNN [13], with an accuracy increase of +1.54 %, precision improvement of +2.66 %, recall improvement of +3.03 %, F1-score improvement of +1.21 %, and a longer execution time.

Table 9 compares the performance of various gene expression analysis methods on the lymphography dataset. The proposed OGT-Net method demonstrates a substantial improvement compared to Fuzzy SVM [12], with an accuracy increase of +73.32 %, precision improvement of +588.88 %, recall improvement of +244.44 %, F1-score improvement of +416.58 %, and a longer execution time. Compared to SMKNN [13], the proposed OGT-Net method shows a significant improvement, with an accuracy increase of +73.32 %, precision improvement of +237.30 %, recall improvement of +203.15 %, F1-score improvement of +227.43 %, and a slightly longer execution time. The proposed OGT-Net method exhibits a notable improvement compared to LMKNN [13], with an accuracy increase of +18.32 %,

Table 4
Ablation study of proposed OGT-Net with Accuracy and Precision Metrics.

Metric	Accuracy (%)				Precision (%)			
Dataset	OGT-Net without LGBM, IHHO	OGT-Net without LGBM feature extraction	OGT-Net without IHHO feature selection	OGT-Net	OGT-Net without LGBM, IHHO	OGT-Net without LGBM feature extraction	OGT-Net without IHHO feature selection	OGT-Net
Prostate	89.43	91.01	93.16	**95.01**	86.88	89.67	91.39	**96.42**
Ovarian	81.05	83.47	85.41	**90.01**	83.17	85.78	87.64	**89.99**
Lymphography	80.74	81.34	82.75	**86.66**	82.61	83.72	84.96	**86.11**
Lung	86.92	89.14	90.87	**91.66**	87.06	89.32	90.21	**91.66**
Colon	85.95	88.29	89.80	**92.30**	83.55	87.44	89.24	**90.0**

Table 5
Ablation study of proposed OGT-Net with Recall and F1-Score Metrics.

Metric	Recall (%)				F1-Score			
Dataset	OGT-Net without LGBM, IHHO	OGT-Net without LGBM feature extraction	OGT-Net without IHHO feature selection	OGT-Net	OGT-Net without LGBM, IHHO	OGT-Net without LGBM feature extraction	OGT-Net without IHHO feature selection	OGT-Net
Prostate	87.45	89.76	91.18	**92.85**	86.72	89.92	91.23	**94.30**
Ovarian	83.11	87.84	89.21	**90.04**	83.67	85.35	87.12	**89.99**
Lymphography	81.45	82.65	84.87	**86.11**	82.01	84.54	85.89	**86.11**
Lung	90.19	90.92	91.12	**92.85**	83.10	88.94	90.01	**91.60**
Colon	91.36	88.67	92.84	**94.4**	85.55	86.20	89.93	**91.50**

Table 6
Overall performance of proposed OGT-Net.

Dataset	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	Execution time (seconds)
Prostate	95.01	96.42	92.85	94.30	2.02
Ovarian	90.01	89.99	90.04	89.99	2.18
Lymphography	86.66	86.11	86.11	86.11	1.97
Lung	91.66	91.66	92.85	91.60	2.344
Colon	92.30	90.0	94.4	91.50	1.004
Average	91.128	90.836	91.25	90.7	1.9036

Table 7
Performance analysis of various methods on colon dataset.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	Execution time (seconds)
Fuzzy SVM [12]	53.84	26.92	50.00	35.34	0.40
SMKNN [13]	61.53	72.27	64.28	57.51	2.83
LMKNN [13]	61.53	72.27	64.28	57.51	3.24
GECC—Net [11]	76.92	81.25	81.25	76.92	0.002
Proposed OGT-Net	92.30	90.0	94.4	91.50	1.004

Table 8
Performance analysis of various methods on lung cancer dataset.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	Execution time (seconds)
Fuzzy SVM [12]	41.66	20.83	50.0	29.41	0.132
SMKNN [13]	90.01	89.91	88.82	89.12	0.03
LMKNN [13]	90.12	89.01	89.81	89.98	0.04
Proposed OGT-Net	91.66	91.66	92.85	91.60	2.344

Table 9
Performance analysis of various methods on lymphography dataset.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	Execution time (seconds)
Fuzzy SVM [12]	50.0	12.5	25.0	16.66	0.88
SMKNN [13]	50.0	25.43	28.33	26.36	0.017
LMKNN [13]	73.33	36.11	40.0	37.87	0.018
GECC—Net [11]	76.66	53.70	55.55	51.95	0.005
Proposed OGT-Net	86.66	86.11	86.11	86.11	1.97

Table 10
Performance analysis of various methods on ovarian cancer dataset.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	Execution time (seconds)
Fuzzy SVM [12]	55.71	27.85	50	35.77	3.84
SMKNN [13]	77.14	76.95	76.50	76.66	0.71
LMKNN [13]	81.42	81.54	80.68	80.95	0.76
Proposed OGT-Net	90.01	89.99	90.04	89.99	2.18

Table 11
Performance analysis on various methods on prostate cancer dataset.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	Execution time (seconds)
Fuzzy SVM [12]	55.0	27.50	50.0	35.48	0.30
SMKNN [13]	70.0	69.69	69.69	69.69	0.015
LMKNN [13]	80.0	81.31	78.78	79.16	0.015
Proposed OGT-Net	95.01	96.42	92.85	94.30	2.02

Table 12
Overall Performance Comparison on various methods.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1-Score	Execution time (seconds)
Fuzzy SVM [12]	51.242	23.12	45	30.532	1.104
SMKNN [13]	69.736	66.85	65.524	63.868	0.7204
LMKNN [13]	77.28	72.048	70.61	69.094	0.8146
Proposed OGT-Net	91.128	90.836	91.25	90.7	1.9036

precision improvement of +137.75 %, recall improvement of +115.27 %, F1-score improvement of +127.16 %, and a slightly longer execution time. Compared to GECC—Net [11], the proposed OGT-Net method showcases a considerable improvement, with an accuracy increase of +13.34 %, precision improvement of +60.35 %, recall improvement of +55.68 %, F1-score improvement of +65.96 %, and a slightly longer execution time.

Table 10 compares the performance of various gene expression analysis methods on the ovarian cancer dataset. The proposed OGT-Net method demonstrates a substantial improvement compared to Fuzzy SVM [12], with an accuracy increase of +61.71 %, precision improvement of +223.60 %, recall improvement of +80.08 %, F1-score improvement of +152.05 %, and a slightly longer execution time. Compared to SMKNN [13], the proposed OGT-Net method shows a

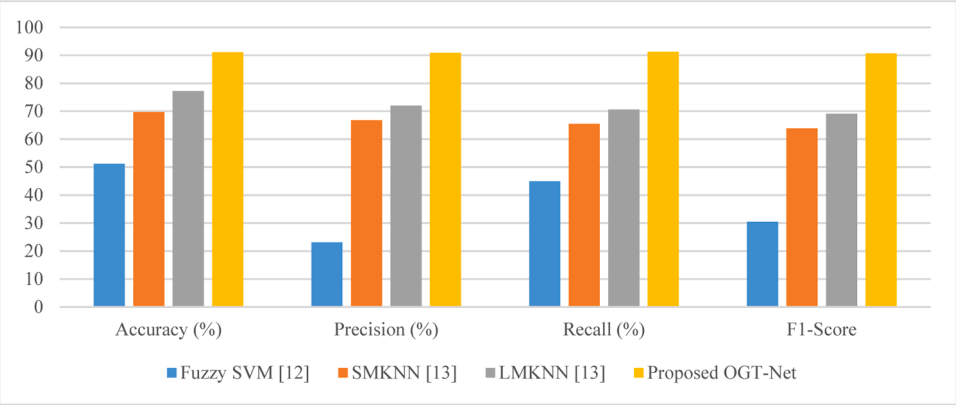

Fig. 6. Overall Performance Comparison of various methods.

significant improvement, with an accuracy increase of +16.55 %, precision improvement of +16.66 %, recall improvement of +17.78 %, F1-score improvement of +17.31 %, and a longer execution time. The proposed OGT-Net method exhibits a notable improvement compared to LMKNN [13], with an accuracy increase of +10.59 %, precision improvement of +10.73 %, recall improvement of +11.97 %, F1-score improvement of +11.49 %, and a longer execution time.

Table 11 compares the performance of various gene expression analysis methods on the prostate cancer dataset. The proposed OGT-Net method demonstrates a substantial improvement compared to Fuzzy SVM [12], with an accuracy increase of +72.74 %, precision improvement of +249.45 %, recall improvement of +85.70 %, F1-score improvement of +165.76 %, and a longer execution time. Compared to SMKNN [13], the proposed OGT-Net method shows a significant improvement, with an accuracy increase of +35.72 %, precision improvement of +38.23 %, recall improvement of +33.68 %, F1-score improvement of +35.49 %, and a longer execution time. The proposed OGT-Net method exhibits a notable improvement compared to LMKNN [13], with an accuracy increase of +18.76 %, precision improvement of +18.71 %, recall improvement of +18.16 %, F1-score improvement of +19.18 %, and a longer execution time.

Discussion

Table 12 and Fig. 6 compare the overall performance of various methods, where the average of all datasets is computed for each metric and method. The proposed OGT-Net demonstrates significant performance improvements over existing methods, namely Fuzzy SVM [12], SMKNN [13], and LMKNN [13]. OGT-Net achieves an impressive accuracy rate of 91.128 %, outperforming Fuzzy SVM, SMKNN, and LMKNN by 39.886 %, 21.392 %, and 13.848 %, respectively. It indicates a substantial enhancement in the model's ability to classify instances compared to the other methods correctly. Looking at precision, OGT-Net achieves a precision rate of 90.836 %, surpassing Fuzzy SVM [12], SMKNN [13], and LMKNN [13] by 67.716 %, 23.986 %, and 18.788 %, respectively. Precision is crucial in assessing the model's ability to avoid false positives, and the proposed OGT-Net excels in providing high-quality, accurate positive predictions. OGT-Net achieves a recall rate of 91.25 %, showcasing superior performance compared to Fuzzy SVM, SMKNN, and LMKNN by 46.25 %, 25.726 %, and 20.64 %, respectively. Recall measures the ability of a model to identify all relevant instances correctly, and the proposed OGT-Net proves highly effective in capturing a significant portion of positive instances. The F1-score, which balances precision and recall, further highlights the excellence of OGT-Net. An F1-score of 90.7 % outperforms Fuzzy SVM, SMKNN, and LMKNN by 60.168 %, 26.368 %, and 21.606 %, respectively. It underscores the proposed model's ability to strike a robust balance between minimizing false positives and negatives.

Conclusion

This work implemented the OGT-Net to identify different types of cancers from gene expression sequences. The dataset was pre-processed by normalizing the records to maintain uniformity. The LGBM extracted correlated features from the pre-processed dataset, and IHHO reduced the number of features by eliminating repeated gene sequences. The customized DLCNN performed the classification of various disorders. The results of the simulations revealed that the proposed OGT-Net performed better than existing approaches considered to be state-of-the-art on a variety of datasets. Compared to existing methods, the proposed OGT-Net method achieves significant accuracy improvement of +20.38 %, precision enhancement of +11.11 %, recall increase of +16.15 %, and F1-score improvement of +18.58 %. Further optimization of the model's architecture and algorithms could lead to even more accurate and robust results. Incorporating advanced technologies, such as ML interpretability tools, could enhance our understanding of the model's decision-making processes and contribute to its transparency. Moreover, the integration of OGT-Net into clinical practice holds immense promise. Developing user-friendly interfaces and tools tailored for healthcare practitioners will be crucial. These tools could empower medical professionals to efficiently leverage gene expression data for cancer diagnosis and individualized treatment planning. Bridging the gap between cutting-edge research and practical, user-friendly applications in clinical settings will be a key focus for future studies.

Funding

No funding received by any government or private concern.

Research involving human participants and/or animals

This article does not contain any studies involving animals performed by any of the authors.

Informed consent

Not applicable.

CRediT authorship contribution statement

Tulasi Raju Nethala: Formal analysis, Data curation. **Bidush Kumar Sahoo:** Formal analysis, Data curation. **Pamidi Srinivasulu:** Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing

financialinterestsor personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data sharing not applicable to this article as no datasets were generated or analyzed during the current study.

Acknowledgements

Not applicable.

References

[1] N. Almugren, H. Alshamlan, A survey on hybrid feature selection methods in microarray gene expression data for cancer classification, IEEE Access 7 (2019) 78533–78548, https://doi.org/10.1109/ACCESS.2019.2922987.
[2] B. He, L. Bergenstråhle, L. Stenbeck, et al., Integrating spatial gene expression and breast tumor morphology via deep learning, Nat. Biomed. Eng. 4 (2020) 827–834, https://doi.org/10.1038/s41551-020-0578-x.
[3] L. Sun, X. Zhang, Y. Qian, J. Xu, S. Zhang, Feature selection using neighborhood entropy-based uncertainty measures for gene expression data classification, Inf. Sci. (Ny) 502 (2019) 18–41, https://doi.org/10.1016/j.ins.2019.05.072. PagesISSN 0020-0255.
[4] Venmathi, A.R, David, S., Govinda, E., Ganapriya, K., Dhanapal, R., Manikandan, A., An Automatic Brain Tumors Detection and Classification Using Deep Convolutional Neural Network with VGG-19, *2023 2nd International Conference on Advancements in Electrical, Electronics, Communication, Computing and Automation (ICAECA)*, Coimbatore, India, 2023, pp. 1-5, doi:10.1109/ICAECA56562.2023.10200949.
[5] H.O. Lee, Y. Hong, H.E. Etlioglu, et al., Lineage-dependent gene expression programs influence the immune landscape of colorectal cancer, Nat. Genet. 52 (2020) 594–603, https://doi.org/10.1038/s41588-020-0636-z.
[6] V.D. Midasala, et al., MFEUsLNet: Skin cancer detection and classification using integrated AI with multilevel feature extraction-based unsupervised learning, Eng. Sci. Technol. an Int. 51 (2024), https://doi.org/10.1016/j.jestch.2024.101632.
[7] B.H. Kim, K. Yu, P.C.W. Lee, Cancer classification of single-cell gene expression data by neural network, Bioinformatics 36 (5) (2020 Mar 1) 1360–1366, https://doi.org/10.1093/bioinformatics/btz772. PMID: 31603465.
[8] A. Manikandan, M. Ponni Bala, Intracardiac mass detection and classification using double convolutional neural network classifier, J. Eng. Res. 11 (2A) (2023) 272–280, https://doi.org/10.36909/jer.12237.
[9] S. Gupta, M.K. Gupta, M. Shabaz, A. Sharma, Deep learning techniques for cancer classification using microarray gene expression data, Front. Physiol. 13 (2022 Sep 30) 952709, https://doi.org/10.3389/fphys.2022.952709. PMID: 36246115; PMCID: PMC9563992.
[10] F. Alharbi, A. Vakanski, Machine learning methods for cancer classification using gene expression data: a review, Bioengineering 10 (2023) 173, https://doi.org/10.3390/bioengineering10020173.
[11] Tulasi Raju Nethala, Bidush Kumar Sahoo, Pamidi Srinivasulu, GECC-Net: gene expression-based cancer classification using hybrid fuzzy ranking network with multi-kernel SVM, in: 2022 International Conference on Industry 4.0 Technology (I4Tech), IEEE, 2022.
[12] Yang Wang, et al., Ensemble-based fuzzy weighted extreme learning machine for gene expression classification, Appl. Intell. 49 (3) (2019) 1161–1171.
[13] S.M. Ayyad, A.I. Saleh, L.M. Labib, Gene expression cancer classification using modified K-Nearest Neighbors technique, Biosystems 176 (2019) 41–51.
[14] M. Monti, J. Fiorentino, E. Milanetti, G. Gosti, G.G. Tartaglia, Prediction of time series gene expression and structural analysis of gene regulatory networks using recurrent neural networks, Entropy 24 (2) (2022) 141, https://doi.org/10.3390/e24020141.
[15] T.-H. Zhang, M.M. Hasib, Y.-C. Chiu, Z.-F. Han, Y.-F. Jin, M. Flores, Y. Chen, Y. Huang, Transformer for gene expression modeling (T-GEM): an interpretable deep learning model for gene expression-based phenotype predictions, Cancers (Basel) 14 (2022) 4763, https://doi.org/10.3390/cancers14194763.
[16] M. Mostavi, Y.-C. Chiu, Y. Huang, Y. Chen, Convolutional neural network models for cancer type prediction based on gene expression, BMC Med. Genom. 13 (2020) 44.
[17] S. Haam, J.-H. Han, H.W. Lee, Y.W. Koh, Cancer Nonimmune-microenvironment-related gene expression signature predicts brain metastasis in lung adenocarcinoma patients after surgery: a machine learning approach using gene expression profiling, Cancers (Basel) 13 (2021) 4468, https://doi.org/10.3390/cancers13174468.
[18] S. Liu, W. Yao, Prediction of lung cancer using gene expression and deep learning with KL divergence gene selection, BMC Bioinformatics 23 (2022) 175, https://doi.org/10.1186/s12859-022-04689-9.
[19] R. T Soto, S. Orozco-Arias, V. Romero-Cano, V. Segovia Bucheli, J.L. Rodríguez-Sotelo, C.F Jiménez-Varón, A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data, PeerJ Comput. Sci. 6 (2020) e270.
[20] L. Ke, M. Li, L. Wang, et al., Improved swarm-optimization-based filter-wrapper gene selection from microarray data for gene expression cancer classification, Pattern Anal. Appl. (2022).
[21] O. Hamzeh, A. Alkhateeb, J. Zheng, et al., Prediction of cancer location in prostate cancer tissue using a machine learning system on gene expression data, BMC Bioinformatic. 21 (Suppl 2) (2020) 78, https://doi.org/10.1186/s12859-020-3345-9.
[22] F. Yuan, L. Lu, Q. Zou, Analysis of gene expression profiles of lung cancer subtypes with machine learning algorithms, Biochimica et Biophysica Acta (BBA) - Mol. Basis Disease 1866 (8) (2020) 165822, https://doi.org/10.1016/j.bbadis.2020.165822. VolumeIssueISSN 0925-4439.
[23] P.G. Díaz, I.S. Berriel, Juan A. Martínez-Rojas, M Ana, Diez-Pascual, Unsupervised feature selection algorithm for multiclass cancer classification of gene expression RNA-Seq data, Genomics 112 (2) (2020) 1916–1925, https://doi.org/10.1016/j.ygeno.2019.11.004. IssuePagesISSN 0888-7543.
[24] J.W. Chen, J. Dhahbi, Lung adenocarcinoma and lung squamous cell carcinoma cancer classification, biomarker identification, and gene expression analysis using overlapping feature selection methods, Sci. Rep. 11 (2021) 13323, https://doi.org/10.1038/s41598-021-92725-8.
[25] M. Divate, A. Tyagi, D.J. Richard, P.A. Prasad, H. Gowda, S.H. Nagaraj, Deep learning-based pan-cancer classification model reveals tissue-of-origin specific gene expression signatures, Cancers (Basel) 14 (2022) 1185, https://doi.org/10.3390/cancers14051185.
[26] A. Yaqoob, N.K. Verma, R.M Aziz, Optimizing gene selection and cancer classification with hybrid sine cosine and cuckoo search algorithm, J. Med. Syst. 48 (2024) 10, https://doi.org/10.1007/s10916-023-02031-1.
[27] B. Li, S. Nabavi, A multimodal graph neural network framework for cancer molecular subtype classification, BMC Bioinformatics 25 (2024) 27, https://doi.org/10.1186/s12859-023-05622-4.
[28] A. Khan, B. Lee, DeepGene Transformer: transformer for the gene expression-based classification of cancer subtypes, Expert Syst. Appl. 226 (2023) 120047, https://doi.org/10.1016/j.eswa.2023.120047.
[29] P.K. Mallick, S.K. Mohapatra, G.S. Chae, et al., Convergent learning–-based model for leukemia classification from gene expression, Pers. Ubiquit. Comput. 27 (2023) 1103–1110, https://doi.org/10.1007/s00779-020-01467-3.
[30] L. Ke, M. Li, L. Wang, et al., Improved swarm-optimization-based filter-wrapper gene selection from microarray data for gene expression tumor classification, Pattern. Anal. Appl. 26 (2023) 455–472, https://doi.org/10.1007/s10044-022-01117-9.

Tulasi Raju Nethala received the M. Tech degree in Computer Science and Engineering from Jawaharlal Nehru Technological University, Kakinada. He is currently a Research Scholar with the Department of Computer Science and Engineering, GIET University, Gunupur, Odisha. His research areas of interest in Data Mining, Machine Learning, Deep Learning. Email: tulasiraju.nethala@giet.edu.

Dr. Bidush Kumar Sahoo received M. Tech and Ph.D. degrees in Computer Science and Engineering from Siksha O Anusandhan University, Bhubaneswar, Odisha. He is currently working as an Associate Professor with the Computer Science and Engineering Department, GIET University, Odisha, India. He has published a number of research papers in international journals (SCI/SCIE/ESCI/Scopus) and conferences. His-research interests include Machine Learning, Cloud Computing, Software Testing. Email: bidushsahoo@giet.edu.

Dr. Pamidi Srinivasulu received Ph.D. degree in Computer Science and Engineering from Acharya Nagarjuna University, Guntur, Completed M. Tech in Computer Science and Engineering from J N T University, Hyderabad. His research areas of interest in Data Mining, Network Security, Machine Learning, Cloud Computing. Email: drspamidi@gmail.com.