Journal of King Saud University – Computer and Information Sciences 35 (2023) 101565

HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

A new approach for cancer prediction based on deep neural learning

Haitham Elwahsh [a], Medhat A. Tawfeek [b,c,*], A.A. Abd El-Aziz [d,e], Mahmood A. Mahmood [d,e], Maazen Alsabaan [f], Engy El-shafeiy [g]

[a] Department of Computer Science, Faculty of Computers and Information, Kafrelsheikh University, Kafrelsheikh, Egypt
[b] Department of Computer Science, College of Computer and Information Sciences, Jouf University, Saudi Arabia
[c] Department of Computer Science, Faculty of Computers and Information, Menoufia University, Egypt
[d] Department of Information Systems, College of Computer and Information Sciences, Jouf University, Saudi Arabia
[e] Department of Information Systems and Technology, Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt
[f] Department of Computer Engineering, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
[g] Department of Computer Science, Faculty of Computers and Artificial Intelligence, University of Sadat City, Sadat City, Egypt

ARTICLE INFO

Article history:
Received 23 February 2023
Revised 16 April 2023
Accepted 19 April 2023
Available online 25 April 2023

Keywords:
DNLC approach
Deep Neural Network learning
Cancer prediction
Cancer classification
Feature selection

ABSTRACT

We know today that numerous factors play a significant role as causes of cancer. Because of this, a doctor's opinion alone cannot be used to classify cancer. Intelligent algorithms providing medical assistance are therefore necessary. In addition, many researchers have adopted them for estimating the likelihood of patient survival, and others have employed predictive methodologies like machine learning and deep learning to forecast prognoses for cancer. The accuracy of predictive cancer prognosis is currently of widespread concern. Since deep neural learning (DNL) methods can quickly predict outcomes from a significant amount of clinical and genetic data, they are essential for predicting various diseases. Deep neural learning is the foundation of our suggested approach. Our deep neural learning cancer prediction model (DNLC) has the following stages. In the first stage, Deep Network (DN) is used to select the best collection of features from datasets. In the second stage, we train genomic or clinical data samples with a deep neural network (DNN). In the third stage, we evaluate the capabilities of the DNLC model of predicting cancer in its earlier stages. For classification, DNLC uses five cancer datasets, which are for colon, lung adenocarcinoma, squamous cell carcinoma, breast, and leukaemia cancers. The five cancer datasets are used in experiments to predict how well the suggested model will perform. The dataset is divided into two parts: training sets, which make up 80% of the dataset, and testing sets, which make up 20%. The experimental results show that the suggested model performs better in terms of accuracy than earlier CNN and RNN models. Our findings demonstrate that the DNLC technique, with an average accuracy of 93%, outperforms other methods in all circumstances.

1. Introduction

Many cell types can develop cancer, which is extremely difficult to identify and diagnose. Tumors, which are abnormal cells, exhibit erratic and random forms. These tumour cells can also be divided into two primary groups: benign and malignant (Aly et al., 2021). At the beginning, the malignant tumour spreads among the nearby tissue cells and hinders the development of healthy tissue cells. The second, in contrast to the first, is a non-cancerous tissue cell and doesn't affect the tissues around it. Even the most skilled pathologists (Khamparia et al., 2021) acknowledged having trouble identifying abnormalities in tissue structure when looking for malignancy. Also, these groups' subtle modifications necessitate particular medical procedures. These could be used with different treatments like surgery, radiation, and medication taken orally.

* Corresponding author.
E-mail addresses: haitham.elwahsh@gmail.com (H. Elwahsh), maelaarg@ju.edu.sa (M.A. Tawfeek), aaeldamarany@ju.edu.sa, a.ahmed@cu.edu.eg (A.A. Abd El-Aziz), mamahmood@ju.edu.sa (M.A. Mahmood), malsabaan@ksu.edu.sa (M. Alsabaan), engy.elshafeiy@gmail.com (E. El-shafeiy).

Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2023.101565
1319-1578/

Many genetic and epigenetic changes interact to cause the complicated disease of cancer (Khachatryan et al., 2016). These changes can influence cell vulnerability, causing the transformation of a normal cell into a cancerous one (Dimitrakopoulos and Beerenwinkel, 2017). According to recent research, malignancies are highly mutated biological entities that primarily disrupt genes involved in cell regeneration, proliferation, and apoptosis, which fuels malignant tumour growth (Ettayebi et al., 2016). With the advent of Next Generation Sequencing (NGS) and the development of genomics, the understanding of cancer biology has been revolutionized, improving diagnosis and therapies, and giving way to the era of precision medicine and the genomics of cancer (Lawrence et al., 2013). Precision medicine is a clinical approach that seeks to select the most appropriate therapeutic method for each patient, considering different clinical parameters and biomarkers. In recent times, systematic studies of the human genome have been carried out to identify recurrent genetic alterations in specific types of cancer, contributing to the understanding of cancer at the molecular level. Single Nucleotide Variations (SNVs), tiny insertions and deletions (Indels), gene fusions, Copy-Number Variations (CNVs), and significant chromosomal rearrangements, also known as Structural Variants (SV), make up the majority of these genetic changes (Cheng et al., 2016). Research has focused on identifying driver mutations, which confer a selective proliferative advantage to cancer cells when compared to normal cells. Such mutations would be causally involved in oncogenesis, the main target for developing new therapies (Helleday et al., 2014). In addition, one of the challenges posed is the differentiation of driver mutations from random or passenger mutations that, until recently, have not been thought to play a significant role in the development of cancer. In addition, regardless of the tumour's histology, these and other mutations may be connected to the cancer diagnosis and prognosis, conferring a predictive value (Helleday et al., 2014). Through these genomic studies, it has been possible to determine that cancer is complex and highly heterogeneous since the genetic mechanisms can vary between patients of the same pathological type (Yang et al., 2014). In this way, it has been possible to identify predisposing genetic variants and tumour subtypes based on the characteristic molecular signatures of each patient (Riazalhosseini and Lathrop, 2016). NGS techniques have allowed researchers to identify genomic variations of human cancers with high mortality and incidence rates, such as lung and breast cancer, through different sequencing techniques such as whole genome (Balloux et al., 2018), whole exome (Whole-Exome Sequencing, WES), and gene subsets of interest (Targeted Panel Sequencing). These advances have facilitated the identification of important altered genetic pathways and provide a genome-wide view of cancers.

Cancer weakens the immune system and produces other cellular changes, which is why it is an ongoing source of concern. The most prevalent malignancies include ovarian, lung, breast, colon, cervical, and prostate cancers. Various previous researchers devised frameworks for determining the likelihood of cancer growth, recurrence, remission, and gauging patient survival (Hayes et al., 1996). For patients, their carers, and clinicians, the accuracy of cancer prognosis is crucial. Accuracy of clinician's predictions is one factor that goes into providing good patient care (Glare et al., 2003). For machines to recognize novel metastatic tumour forms or to diagnose a disease at an early stage when treatment is more difficult, they need to employ disease detection methods that can classify tumour types and identify cancer symptoms. Over the past three decades, prostate, lung, and breast cancer have been the most likely sex-specific cancers to develop in men and women. Therefore, it was crucial to create a reliable early-stage cancer forecasting model. Support Vector Machine (SVM) is widely used in various fields, including medicine, and gives higher prediction performance in both linear and nonlinear circumstances. Cancer predictive models are still in their infancy, despite SVM being a remarkable classifier (Statnikov et al., 2005). Breast cancer classification, segmentation, and detection have been recently addressed using artificial neural network (ANN)-based approaches (Cheng et al., 2016) (e.g., external neural networks (SNNs) (Nhu et al., 2020) and deep neural networks (DNNs) (Murtaza et al., 2020), as well as machine learning (ML) techniques like support vector machines, logistic regression, naive Bayes classifiers, and decision trees. Deep neural networks (DNNs) frequently contain two or more hidden layers between the input and output, in contrast to numerous shallow neural networks (SNNs). Only a small number of publications (Murtaza et al., 2020; Deshmukh and Kashyap, 2022; Houssein et al., 2021) describe the classification of breast cancer in medical imaging modalities. Other studies (Dar et al., 2022; Travis et al., 2013) looked at the advantages of applying hand-engineered features against using ML techniques to analyse breast cancer images. The mutation test (Pao and Ladanyi, 2007) has emerged as a crucial tool for selecting the most effective patient treatments in clinical trials. For unidentified mutations, direct sequencing is a screening-based substitute for indirect sequencing. The mutation test for Epidermal Growth Factor Receptors (EGFR) has been identified as a genetic mutation test for lung cancer (Cong et al., 2020). ANN and SVM, two types of classification tools, contrast with their non-ensemble counterparts. Deep learning is a subset of artificial intelligence included in the same category. Deep learning is a technique for extracting features, such as text, images, or speech, from data. Deep learning is one of AI's most important features (xxxx). Traditional AI systems require many steps to complete image recognition tasks, including preprocessing, feature extraction, careful feature selection, learning, and evaluation (xxxx). The introduction of these systems relies on the chosen features, which may or may not be the best attributes for distinguishing between classes. In contrast, deep learning, uses automated learning of functions for various tasks rather than the conventional AI approach. It can learn and process datasets in a single step (Pohekar and Ramachandran, 2004).

Metaheuristic algorithms fall into two categories: single-solution and population-based. In the latter type an optimization algorithm uses an initial random search agent that represents the population. In contrast, in the former type, an optimization algorithm performs the optimization process with a single candidate solution that changes and is updated between iterations. Each answer to the optimization problem becomes the new candidate for the next search agent. The agents cooperate and exchange knowledge about the search region to prevent local optimization from stagnating and to achieve their universal aim. Numerous studies (Mela et al., 2012; Xu and Zhang, 2014) have used different optimization strategies to address decision-making challenges. Metaheuristic algorithms must maintain proper control and a healthy balance between exploitation and exploration (Mafarja and Mirjalili, 2017). Exploitation is the ability to identify superior solutions to accepted theories. Exploration involves using metaheuristics to seek a larger search space for new sites. Most metaheuristics use exploration early in the optimization process to thoroughly assess the feasible region and avoid a recession in the local optima. Several metaheuristic techniques have been combined with wrapper methods to provide a satisfying result in a reasonable period.

In addition to using a single optimization technique to address the Feature Selection (FS) problem, researchers have developed other hybrid approaches to handle binary optimization challenges. For example, the Whale Optimization Algorithm (WOA) and simulated annealing hybrid strategy are investigated in (xxxx), and the Genetic Algorithm (GA) and Particle Swarm Optimizer (PSO) hybrid technique are reported in (Shukla et al., 2019). A hybrid technique that combines the filter and wrapper approaches of FS

has already been researched (Tubishat et al., 2021). However, there is no guarantee that the FS problem will yield a better selection of traits. Furthermore, the No Free Lunch (NFL) theorem prohibits using any optimizer to solve optimization problems (Alzahrani and Venneri, 2015), which explains why some optimizers do not perform well when dealing with specific optimization challenges.

The importance of early cancer tissue detection cannot be overstated. Hence, it takes a lot of time and effort for pathologists to recognise tumour patterns from the visual perception of cancer tissue data (Meuten et al., 2021). It is urgently necessary to create an automated computer-aided method to aid pathology specialists in recognising cancer. Analysis of histopathology images by pathologists will take less time and effort (Zováthi et al., 2022; Madabhushi et al., 2011). In order to identify malignancy, ML/DL is frequently employed. However, the majority of researchers rely on just one deep learning model, such as RNN, LSTM, CNN, etc. As a result, it was determined that the performance of these models was inadequate. Hybrid DL models are always capable of enhancing classification performance (Kanna and Santhi, 2022). This research provided a DNLC strategy to enhance the classification performance and efficiency identify cancer tissues in order to address the aforementioned significant concerns.

The research's principal contributions are listed below. In order to lessen the pathologist's inaccuracy, a new Deep Neural Learning model is provided in this study that extracts fewer features from histopathology data and classifies them. It is suggested to classify cancer detection in clinical research effectively using the DNLC technique. We have compared the major performance measures (Acc (%), accuracy (%), sensitivity (%), and specificity (%), F1-score, and AUC, with the present ML/DL model applied on the five datasets as part of the evaluation process for the proposed DNLC approach. to determine how well the hybrid models perform in terms of classification. Comparing the proposed hybrid model to previous hybrid DL models, it is discovered that its classification results are excellent.

The paper is structured in the following manner. A review of the literature and information pertinent to the investigation are found in Section 2. Deep neural network details are provided in Section 3, and the proposed DNLC method is described in Section 4. The experimental results are covered in Section 5. A summary of the conclusions and suggestions for additional research are included in the conclusion of Section 6.

2. Related works

Given the complexity of genomic data from NGS in terms of volume and complexity, new architectural and computational tools are required to enable the deployment of genomic analysis in clinical practice. Big data is the concept that has recently been used to fill this demand. This idea encompasses a variety of activities, such as gathering, processing, and analysing large amounts of data from many sources to characterize the data, identify patterns and correlations among them, and forecast specific therapeutic responses that could be of potential interest. There are many different techniques used to carry out these steps, including data mining. These techniques seek to convert raw data into assets of great value based on hidden patterns in the data which identify useful information (Loyola-González et al., 2013). One approach that allows the identification of hidden patterns is machine learning (ML), which involves the analytical methods of classification, clustering, and linear regression, with emphasis on predicting variables of interest. These algorithms fall into three categories: unsupervised learning, semi-supervised learning, and supervised learning. Outputs can be predicted via supervised learning, which uses training data that corresponds to already classified or labelled information. As an illustration, consider classification approaches, which create rules for categorizing items into groups based on a vector of measurements on these objects (also known as a predictive variable). Examples of classification techniques include logistic regression (LR), naïve Bayesian (NB), decision trees (DT), neural networks (NN), Bayesian networks, and support vector machines (SVM) among others (Gonzalez-Ericsson et al., 2020). Unsupervised learning tries to find patterns within the data without information about the output that groups the data. One example of this technique is clustering. This approach is used to find clusters in data using distance metrics. There are different clustering techniques, such as *k*-means and principal components-based clustering. These methods have been extensively employed in phylogenetic research, microarray analysis, and most recently the study of complex disorders. The goal of semi-supervised learning is to balance accuracy and performance. Recent studies have raised the possibility of applying these ML techniques to evaluate, and classify tumour types, predict clinical responses in different diseases, and identify mutational patterns associated with clinical phenotypes, among other things (Wu et al., 2019). The main objective is to determine the factors most relevant in the appearance and progression of cancer, thus contributing to the selection of the best therapeutic strategy. Researchers are now using ML to unearth hidden insights within their data to resolve classification or prediction challenges. The aim of cancer research is to find the best course of treatment. To this end, researchers have begun using ML models in their studies to classify or predict key clinical outcomes, including overall survival (OS) and distant metastasis (DM) (Chong et al., 2021). These models provide crucial insights that help in the selection of therapeutic approaches. To predict overall survival rates, researchers have turned to investigating MRNA expression data, somatic mutational features, mutation of driver genes, and other factors to extract predictive value from this information. However, the variables driving clinical progression in distant metastases are often unknown or are understudied, in contrast to survival research, which directly affects clinical and survival prognosis.

The researchers in (Polat and Güneş, 2007) employed ML to find lymph node metastases from lung adenocarcinomas (LUAD) based on DNA methylation traits. Despite recent advancements, further research is needed in this area to improve the accuracy of diagnosis and prognosis for lung cancer patients, and to better comprehend the mechanism for metastasis growth, since metastases are the leading cause of morbidity and mortality in cancer patients. Identifying the most significant clinical and genetic characteristics that contribute to metastasis is crucial for early detection of cancer. With the help of this information, medical staff might better be able to choose the most effective treatment for LUAD patients. However, this area of study is quite challenging, because of the limited number of datasets that contain information on cancer patients' metastatic stages, which would be of great medical relevance. Furthermore, the significant differences between individuals with metastasis and those without are problematic for precision medicine. Merely possessing the knowledge of these differences is insufficient to assist with solving such problems. Lung cancer was identified by Polat et al. (Selvanambi et al., 2020) using fuzzy weighting pre-processing, principal component analysis (PCA), and a counterfeit-resistant acknowledgment system. The framework is divided into three phases. By a standard part examination, the dataset is constrained to four main highlights with 57 sub highlights. Second, before the basic classifier, a weighting plan based on fluffy weighting pre-handling was implemented as a preprocessing step. Third, a classifier was created using a counterfeit-safe recognition method. Tests were run on the lung dataset to analyse tumours entirely automatically. The fact that the framework's characterization precision was 100% for upcoming grouping applications was very promising. Fang et al. sought to identify and

validate characteristics related to lung cancer growth and associated pathways. A network-based biomarker identification method with quality set improvement evaluation was introduced in (Laxminarayanamma et al., 2022). In addition to the qualities anticipated by earlier results in research into cigarette smoking and lung cancer, researchers also found many novel and surprising traits with potential physiological effects from cigarette smoking or lung cancer growth using the linked revelation approach. Thus, it was shown that disease-specific network biomarkers, interactions between traits or proteins, or cross-talking pathways contribute more specifically to promoting precise lung treatments. Krishnaiah et al. (Naik and Edla, 2021) have applied this conclusion on various lung tumour datasets employing information mining approaches and streamlining methodologies to examine lung tumours. Knowledge discovery in databases (KDD) involves applying data mining techniques to identify and utilize patterns of malignancy across a wide range of characteristics, as well as to predict the course of an illness using specific therapy instances contained in datasets. In their study, Kuruvilla and colleagues (Sampedro et al., 2014) used a computer-aided diagnosis (CAD) technique for a neural network-based tumour recognition in computed tomography (CT) pictures. This involved extracting the entire lung from the CT images and identifying features of the dispersed image. Various statistical measures such as mean, SD, skew, kurtosis, and the fifth and sixth central moments were calculated to characterize the tumours. The characterization features were then fed into a feed-forward/backpropagation neural network for classification. Chauhan et al. (Senthil and Ayshwarya, 2018) studied ANN (artificial neural network) methods, such as picture processing, LDA (linear dependency analysis), and SOM (self-organizing maps), for the purpose of detecting lung tumours. In conclusion, they advised using support vector machines as characterization tools. SVM are learning models that analyse data to find patterns. Dcruz et al. (Shaffie et al., 2018) initially devised a method for detecting lung tumours. Information pre-processing is carried out first in this method to improve the image. The datasets are then generated and tested using information mining and neural networks, which are essential for differentiating potential treatments. Using a back propagation neural network (BPNN) to classify the information images into harmful and non-threatening categories led them to their conclusion. At the start of the process, the extracted feature photographs determine the stage of malignancy, which helps the doctors make a diagnosis. A plan for predicting lung disease, as well as early diagnosis and treatment of lung tumours, has been put forward by Taher et al. (Qian et al., 2019). Different features were extracted from the photos to prepare for lung disease classification. They found that design acknowledgment-based systems are crucial for predicting lung tumours. Based on their prior knowledge of picture-handling procedures, they presented a thorough study of the causes of lung tumours. The coupling of picture preparation techniques with computational understanding-based technologies is beneficial for the prediction and fundamental management of lung tumours.

Qian et al. (Abdeldayem and Bourlai, 2018) use a hybrid information fusion technique to present an emotion-aware recommendation system that analyses user ratings as explicit information, user social network data as implicit information, and sentiment from user reviews as emotional information. This approach produces prediction ratings and suggestions that are more precise. Zhang et al. (Huang et al., 2015) proposed a hybrid approach incorporating fiducial and non-fiducial characteristics to extract more specific electrocardiogram (ECG) features and improve authentication stability. They developed a parallel ECG pattern recognition framework to boost the effectiveness of recognition in various ECG feature spaces. Experiments were conducted to validate the performance of the proposed authentication method. Back propagation neural networks were employed by Xiao et al. (Huang et al., 2017) to develop a price-predicting model. They then proposed a self-evolving commodity futures trading method based on futures market regulations. The approach is back-tested using data from the Shanghai Futures Exchange and the Dalian Futures Exchange. A comparison is made between their proposal and conventional tactics to demonstrate how they differ. Experiments show that their strategies perform better than the other approaches they compare in the evaluation. Their strategy performs better than the market in terms of yield and risk. According to Zhang et al. (Ozturk and Unal, 2020), current software-defined network (SDN) applications in the vehicular network mainly concentrate on data communications between vehicles and other devices. In contrast, the vehicular controller area network is still restricted to a few specific applications, only offering users basic services and unable to meet the demands of a complex driving environment. They offered an SDN-based approach for creating a safety-oriented vehicular controller area network that can guarantee traffic safety through driver fatigue detection and emotional recognition, monitored through the driver's physiological and psychological status.

Small-cell lung cancer (SCLC) is one type of cancer that urgently requires a novel drug and is a malignancy without a well-defined treatment that often spreads early. Chemotherapy, radiation, and surgery help only 6% of patients live five years following diagnosis. Between 10 and 15 percent of lung cancers are SCLC. A model or framework for predicting SCLC needs to be developed. Detecting false positives in lung cancer CT imaging is another unresolved problem. Another challenging task is finding the dataset's most important properties for classification, grouping, or prediction.

Effective studies have been conducted on the treatment of Parkinson's disease, as it causes movement disorders. An effective classification method has been proposed for data from Parkinson's disease and normal individuals. Since the data set used in the proposed study consisted of duplicate samples, it was difficult to use independence-based classifiers for the duplicate data set. The distribution of features in the data set was examined, and the success of traditional classifiers was very low due to the fact that the distribution centers between the groups are very close to each other.

Based on the basic idea that a greater distance of centers of mass from each other will increase success, dimensional techniques such as PCA, ICA, Relieff and RICA were used. When the desired success was not achieved, the bond theory was generated using a two-stage whale optimization algorithm. The features of the three samples taken from an individual were closed to each other in the feature space, the aggregate samples belonging to the same class were drawn on one side of the feature space and the feature space of the other class was placed in the furthest position from the center point.

Thus (Öztürk et al., 2018), the different samples belonging to classified the Parkinson's disease Voice Recordings. The proposed method was compared with other techniques and the result was that the representation ability in feature space is stronger than other related methods.

The study (Özkaya et al., 2021) the significance ratios of the features extracted from the images were examined using feature extraction algorithms. Significance coefficient was determined for each feature parameter. The number of features is reduced according to the importance weight calculated for each feature. Classification success was examined for each case. Six feature extraction algorithms were used for this. The classification success of all these feature extraction algorithms was examined separately. Then, all the properties were combined to form a single property matrix. The obtained characteristic matrix was reduced using principal component analysis and dilution methods.

This study (Öztürk and Çukur, 2022) achieved an improvement in design dimensions for multi-scale applications by using geometric shapes with higher ability to reflect shapes. This aims to optimize some key dimensions using deep learning models in order to obtain shorter performance optimizations. As a result of these models, the designs were re-evaluated in LSTM-1 + Dropout layer-1 + LSTM-2 + Dropout layer-2 simulation software representing an advanced deep learning model. It was found that traditional methods are not sufficient to improve this problem. This problem was solved using the deep learning model. In particular, the performance of this deep learning model reduces the optimization time in the model used.

Melanoma is a potentially fatal, treatable skin cancer that greatly increases the survival rate when diagnosed in its early stages (Sharma et al., 2016). Learning-based methods are very promising for the detection of melanoma from stereoscopic images. However, because melanoma is a rare disease, existing databases of skin lesions often contain very unbalanced numbers of benign versus malignant specimens. In contrast, this imbalance introduced a significant bias into the classification models due to the statistical dominance of the majority class. To solve this problem, they introduced a deep synthesis approach based on the inclusion of latent space for stereoscopic images. Clustering was achieved using an innovative margin-free triplet loss (COM-Triplet) superimposed on images grown from a convolutional neural network backbone. Aiming at maximally separating cluster centers rather than minimizing classification error, this proposed method was less sensitive to class imbalance. To avoid the need for labeled data, a COM-Triplet was implemented based on pseudo-labels generated by a Gaussian mixture model (GMM). Extensive experiments showed that deep clustering with loss COM-Triplet outperformed clustering with triplet loss, competing classifiers in both supervised and unsupervised settings.

In this study, we present a method to predict the existence of metastasis in cancer sample datasets using a new approach based on a deep neural network classifier model. The predictors used in this method include mutational load and pertinent clinical parameters. We could choose an appropriate processing, training, and validation procedures for a deep neural learning model by considering the research data's specific features. This benchmarking will make it possible to understand better the connections between genetic and clinical markers for cancer prognosis.

3. Methodologies

Several cancer diagnostic studies have used a variety of methodologies for cancer diagnosis prediction, with some of these methods demonstrating high prediction accuracies. Several researchers employ ML classifiers to improve treatment and medicine discovery for diagnosis using K-nearest neighbour (KNN), LR, DT, random forest (RF), and SVM. In (van Vliet et al., 2008), a cervical cancer dataset was used. Furthermore, (Andersen et al., 2004) separately employed four different breast cancer datasets. Similarly, in (Kang et al., 2022), two separate colon cancer datasets were used. A multi-layered DL algorithm was used in (Mohammed et al., 2021) to recognize the type of disease in diverse microarray data.

To classify microarray data, (Withnell et al., 2021) suggested an ensemble of ANN classifiers. They employed four different cancer datasets in their research. Combining gene FS with cancer classification for gene expressions and other types of omics data is recommended in (Zamry et al., 2021). Meanwhile, Adiwijaya (Mohd Amiruddin et al., 2020) used Principal Component Analysis to reduce dimensions on SVM and Local Mean Binary Pattern (LMBP), and PCA was used in (Buchman et al., 2022) to work on ANN and GA.

We chose DL over other traditional methods because we need to be able to handle large data sets. Furthermore, the time taken to produce a cancer diagnosis is essential because the patient's life depends on it, especially in serious circumstances. In this scenario, DL is the most appropriate methodology to employ because with high-end infrastructure it can be trained in a reasonable amount of time. Recurrent neural networks (RNNs) are a form of Neural Network (NN) that deal with serial data input and output.

With feedback to the neural networks' feed-forward (FF), RNNs record the temporal relationship between input/output sequences. RNNs help with speech recognition when dealing with sequential data.

3.1. Recurrent neural network (RNN)

Let X = $\{\boldsymbol{x_t}\}$ be the input to an RNN, where $\boldsymbol{x_t} \in \boldsymbol{R^N}$ is an input vector for each time step t. Consider the output as Y = $\{\boldsymbol{y_t}\}$, where $\boldsymbol{y_t} \in \boldsymbol{R^M}$ is the vector that represents the output for each time step t. Our goal is to model the P (Y|X) distribution. The following, by Selvanambi (Mohammed et al., 2021), determines the output of the RNN $\boldsymbol{y_t}$:

$$\boldsymbol{P}(\boldsymbol{y_t}|\{\boldsymbol{x_i}\}_{i=1}^{t}) = \boldsymbol{\sigma}(\boldsymbol{W_y h_t} + \boldsymbol{b_y}) \quad (1)$$

Where:

$$\boldsymbol{h_t} = \boldsymbol{tanh}(\boldsymbol{W_h h_{t-1}} + \boldsymbol{W_x x_t} + \boldsymbol{b_h}) \quad (2)$$

$\boldsymbol{W_y}$denotes the weight matrix that connects the hidden layer to the output layer (y). $\boldsymbol{W_h}$ denotes the weight matrix that connects the hidden layer to the hidden layer (h).

$\boldsymbol{W_x}$ denotes the weight matrix that connects the input layer to the hidden layer (x). The hidden layer bias vectors represent the output layer's bias vectors, which are denoted by $\boldsymbol{b_h}$, and the output layer bias vectors are denoted by $\boldsymbol{b_y}$.

The sigmoid σ, tanh, and rectified linear unit (ReLU) activation functions indicate final nonlinearity in classification. The recurrent network computes the output $\boldsymbol{y_t}$ based on the information propagated from the hidden layer in every situation where it depends directly or indirectly on the values $\{\boldsymbol{x_i}\}_{i=1}^{t} = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x_t}\}$as shown in Fig. 1.

Bidirectional RNNs (BRNNs) offer the advantage of passing more input information to the network. Fixed input data and the absence of future information plague multilayer perceptron networks (MLPs) and time delay recurrent networks (TDNs). However, when there is no fixed input data and future information is inaccessible, BRNNs solve one of these concerns.

3.2. Convolution neural network (CNN)

CNNs are one of the most popular DNN architectures. CNNs mimic both human and animal brain activity [66]. CNNs engage in various initiatives, including categorization, pattern, identification, etc. CNNs contain a set of connected layers; each layer performs a specific task. CNNs generally contain three layers: convolution, nonlinear, and pooling. The convolution layer is the central component of the model and the most crucial of the three layers. Without altering the amount of inputted data, the convolution layer accepts the data and recreates a map of the data. The primary objectives of the convolution layer are to extract various characteristics, distribute the weights of features from preceding layers, and produce good results with the fewest possible parameters. The nonlinear layer or modelling layer are two names for the second layer. The nonlinear layer, which is employed in the modelling process, employs a nonlinear activation function. The statistical pooling layer is the third type used to reduce the number of

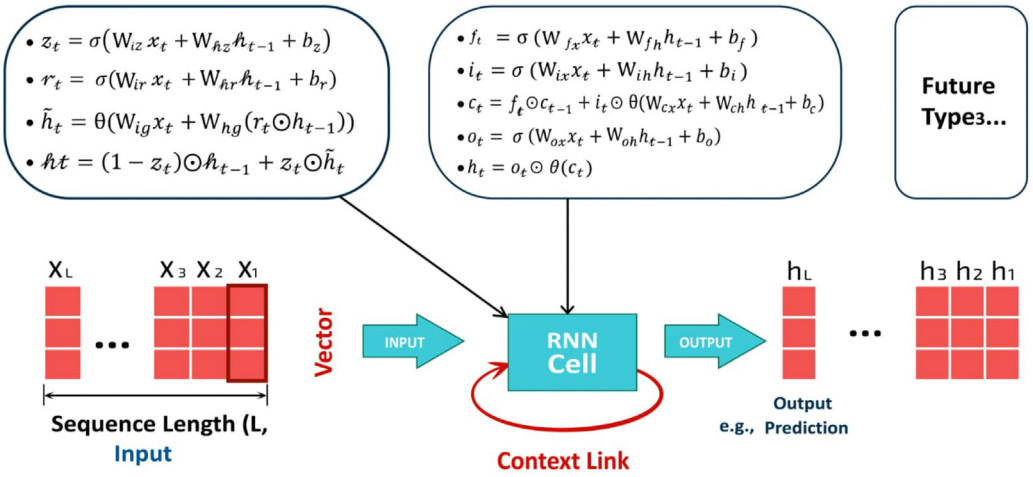

Fig. 1. Multiple RNN cell types.

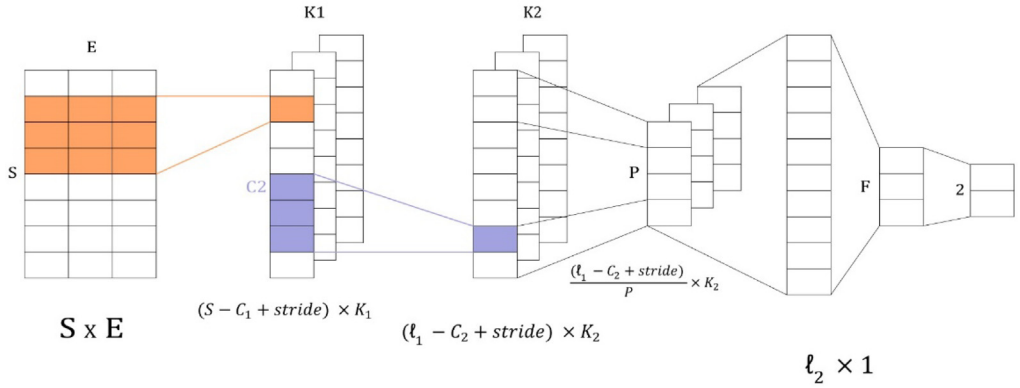

Fig. 2. Architecture of convolution neural network.

dimensions and to carry out statistical operations like mean and max (Fig. 2).

4. The proposed DNLC approach

As stated by Khodayar and Teshnehlab (Zamry et al., 2021), the autoencoder is regarded as a nonlinear method suitable for severely nonlinear datasets. Fig. 3 shows the method known as the DNLC approach. It shows how the PCA method reduces the number of features from **X to $\widehat{\mathbf{X}}$.** The DNN then creates the X′ features by removing features from the X features. After that, the X′ features are categorized using a classifier (deep network). According to (Mohd Amiruddin et al., 2020), pre-processing techniques are required because some datasets include problems like data imbalance or datum shift.

The k-dimensional subspace with the maximum variance in the data is found through PCA. For k = 1, given a data matrix X R n p with a mean of 0 for each column, one aims to obtain a minimum number of features that nevertheless capture most of the variance using PCA.

For instance, it would be preferable for the major components of a gene cancer dataset to include just a few important genes, making it simple for humans to analyse. Therefore, it is necessary to enforce the PCA component's sparsity, which results in a trade-off between explained variance and sparsity.

With [0, 1] as the regulating parameter for sparsity. For values equal to 0, standard PCA is recovered; however, for values equal to 1, the component with the highest variance is the sparsest non-trivial solution. The formulation's fit inside our broad framework is immediately obvious (we use this notation: $\widehat{\mathbf{X}}$ = max(0, x) in the following).

The proposed method can be applied irrespective of how a dataset is balanced or unbalanced. Hence, those pre-processing steps are not required in this research. As a result, computing will become less complicated. However, choosing the appropriate structure for DNLC is crucial. We used the same framework for all datasets (Fig. 3). Fig. 3 displays the dimension (W) of the weight matrix, or the number of neurons in the hidden layer that are adaptive, which means the number of hidden neurons are 2/3 the size of the input layer plus the size of the output layer. Additionally, these statistics show the number of features gathered from each autoencoder layer. Since not all datasets are nonlinear, increasing the number of features in the first encoder layer is no longer needed.

Each time, we first consider the number of autoencoders and the number of retrieved features while taking into account the number of neurons in the autoencoder layer. Next, the error is calculated for each of these configurations. We then select which number of neurons in the autoencoder produces the smallest error. In the final step we employ the various classifiers to classify the features. Equation (5) must be considered when training the autoencoders in this method. Equations (3) - (4) are the forward

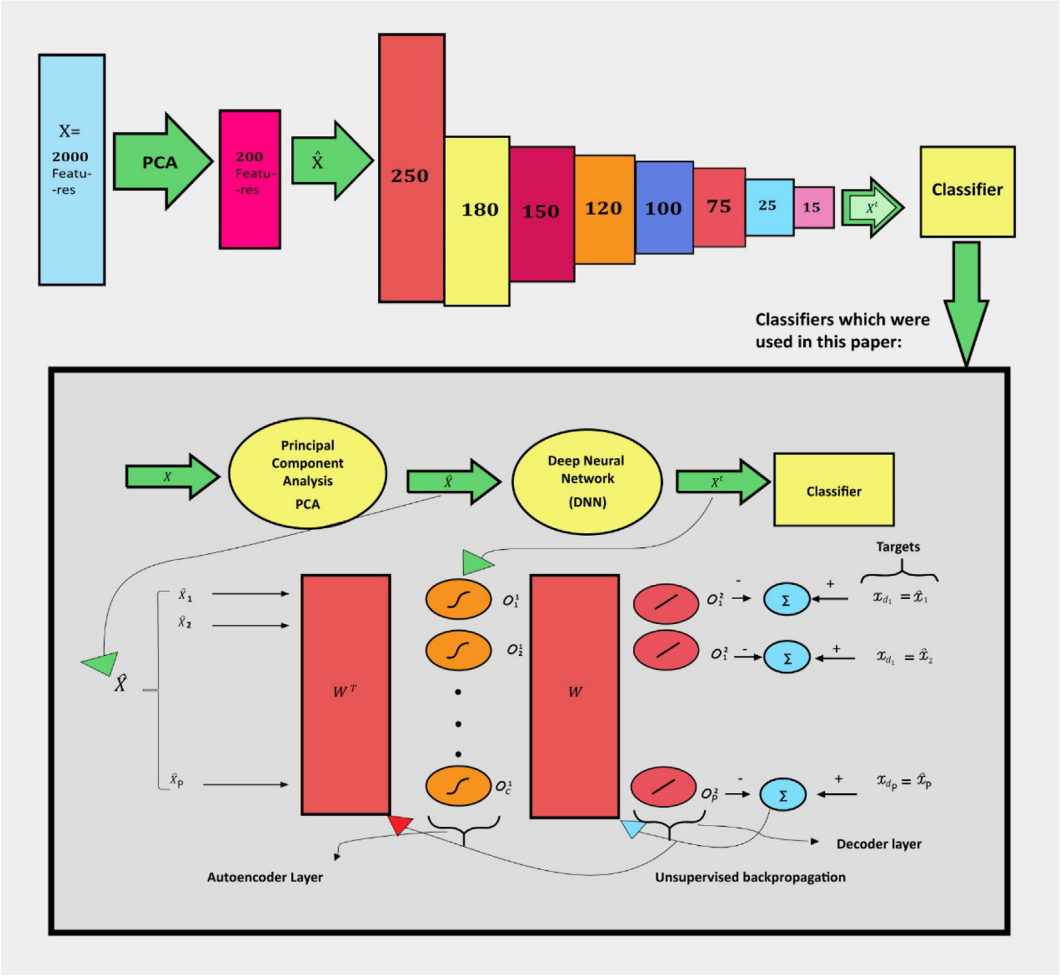

Fig. 3. Structures in the DNLC approach.

neural network equations. The activation functions in the autoencoder and decoder layers are given in Equations (5) and (7) as linear and tangent sigmoid functions. In this case, we apply the gradient descent optimization technique to update the weight matrix connecting the neurons in the hidden layer to the output layer W^2. In gradient descent, each encoder's output should be compared to its aim, as shown in Fig. 3. The inputs to the targets and the encoder layer are the same (unsupervised learning). We then use Equation (8) to calculate the error, where the term j indicates the j^{th} error. As a result, the cost function is described as a sum square error (9).

Moreover, to increase the training phase's effectiveness. The cost function for this learning process is represented using Equation (10), where the coefficients k_1 and k_2 signify the effect of recent and historical errors. Here, considerations of 1 and 0.5 were given to variables k_1 and k_2, respectively. After that, the estimated error will be sent back to update W^2 using a chain rule: Equations (11) and (12). After updating W^2, we transpose this matrix and substitute W^1—which connects input layer neurons with hidden layer neurons—for W^2.

The equations depicted below are the initial steps of the feed-forward algorithm.

$$W^2 = W; W^1 = W^T \tag{3}$$

$$net^1(k) = W^T(k) + \hat{X}(k) \tag{4}$$

The following equation is used to determine the outputs of the auto-encoder layers:

$$O^1(k) = \text{tansig}(net^1(k)) \tag{5}$$

$$net^2(k) = W(k)O^1(k) \tag{6}$$

As seen in the equation below, a linear function determines the outputs of the decoder layers:

$$O^2(k) = net^2(k) \tag{7}$$

The learning phase, illustrated by (8)–(11) and (12), is the second step.

$$e_j(k) = X_{d_j}(k) - O^2(k) = X_{d_j}(k) - \left(W(k)O^1(k)\right) \tag{8}$$

Additionally, cost functions (9) and (10) are considered.

$$E_1(k) = \frac{1}{2}\sum_{j=1}^{p} e_j^2(k) \tag{9}$$

$$\begin{aligned} E_2(k) &= \tfrac{1}{2}\textstyle\sum_{j=1}^{p} r_j^2(k) = \tfrac{1}{2}\sum_{j=1}^{p}\left(k_1 e_j(k) + k_2 \dot{e}_j(k)\right)^2 \\ &= \tfrac{1}{2}\textstyle\sum_{j=1}^{p}\left((k_1 + k_2)e_j(k) - k_2 e_j(k-1)\right)^2 \end{aligned} \tag{10}$$

Traditional backpropagation is described in (11).

$$\Delta W(k) = -\eta \frac{\partial E_1(k)}{\partial W(k)} = -\frac{\partial E_1(k)}{\partial e(k)} \frac{\partial e(k)}{\partial O^2(k)} \frac{\partial O^2(k)}{\partial W(k)} = -\eta(e(k))(-1)O^1(k) = \eta e(k) O^1(k) \quad (11)$$

Where η is the learning value that should be $0 < \eta \leq 1$ for keeping the learning process stable.

$$\Delta W(k) = -\eta \frac{\partial E_2(k)}{\partial W(k)} = -\frac{\partial E_2(k)}{\partial r(k)} \frac{\partial r(k)}{\partial e(k)} \frac{\partial e(k)}{\partial O^2(k)} \frac{\partial O^2(k)}{\partial W(k)} = \eta(k_1 + k_2) r(k) O^1(k) \quad (12)$$

Extracted features will be considered as the input for each classifier after updating weight matrices in each autoencoder. We separate the samples into two categories for each classifier: the training and testing samples (unseen data). Since the exact categorization for each dataset is known, we utilize a supervised learning technique to categorize the features in this case.

The DNLC approach suggests a structure for classifying five cancer datasets to improve the accuracy of the results. This proposed approach had never been used with these five cancer datasets before.

5. Experimental DNLC approach

The focus of the present study was cancer detection. The prediction was performed using the DNLC Approach, while the detection employed five distinct datasets. The CNN was executed in order to obtain state-of-the-art architectures (SOTA) performance metrics. Throughout the various experiments, achieving high-performance metrics was the objective. The length of time required for learning and processing was lengthy, so it was not specified in this study. It is important to note that the duration largely depended on the working environment. The current research utilised Google Colab. The dataset is divided into two parts: training sets, which make up 80% of the dataset, and testing sets, which make up 20%.

Five cancer datasets are available for breast, leukaemia, squamous cell carcinoma, colon, and lung cancers (Buchman et al., 2022). Gene expression and clinical data are present in all five datasets. Using these five datasets, we evaluated our suggested strategy and technique. The traits in the five cancer datasets we chose are noticeably varied. The colon cancer dataset's features come in a wide variety, and the total feature value ranges from 5.8163 to 20,903. This dataset is frequently nonlinear.

The number of samples and each dataset's properties are shown in Table 1. As is evident, these five datasets include a wide range of values for the number of features. In this case, the number of hidden layers will increase if we utilize a traditional neural network. Weight matrices in the initially hidden layers close to the input layer will not be changed. Because of this, a traditional neural network cannot be used. Therefore, this suggests the need for feature extraction and reduction, which are carried out using the PCA and DNLC approaches.

5.1. Experiments evaluation metrics

Many evaluation matrices, such as the confusion matrix, which frequently provide different classification metrics and performance evaluation parameters, can be used to assess the performance of the trained model. True Positive (TP) refers to the quantity of accurate positive predictions. The number of inaccurately pessimistic projections is known as the False Negative (FN) rate. The term False Positive (FP) refers to the quantity of incorrectly positive predictions. True Negative (TN) is the total amount of valid negative predictions. The equations below display the details of the evaluation measures obtained from the confusion matrix.

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (13)$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (14)$$

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (15)$$

$$P - \text{value} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (16)$$

$$N - \text{value} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (17)$$

5.2. Lung adenocarcinoma (Squamous cell carcinoma dataset)

Even for the experienced eye, distinguishing between squamous cell carcinoma and other lung cancer subtypes can be difficult and time-consuming. Clinical and somatic mutational data from the Pan-Lung Cancer 2016 dataset were made available by the Cancer Genome Atlas (TCGA) (Campbell et al. 2016). This dataset contains 1144 lung adenocarcinoma and lung squamous cell carcinoma (LSCC) patients, which is a significant number of samples to produce accurate findings for deep learning. The Pan-Lung Cancer 2016 dataset is one of the few that include details on the patients' metastatic status, which is the variable of relevance in this study. We separated the patients into two age groups: up to and including 60, and older than 60. Table 2 displays the characteristics of the entire cohort.

In this paper, Python was used to implement performance analysis in the Keras tool. With a learning rate of 0.001, a batch size of 40, and 150 epoch values, the Adam and SGD optimizers were utilized. To measure the performance of LeNet and AlexNet, various optimizers such as Adam and SGD were implemented, and various statistical parameters were measured to detect cancer.

And using the OpenSlide library, the slides were tiled in non-overlapping 512x512 pixel panes at a magnification of x20 (533 of the 2167 slides initially uploaded were removed because of compatibility and readability issues at this stage). The low-information slides—all the tiles with a background covering more than 50% of the surface area—were eliminated (where all the values are below 220 in the RGB colour space). In the procedure, almost 1,000,000 tiles were produced.

Using DNA methylation characteristics, we employed machine learning to differentiate lymph node metastases from LUAD ones.

Table 1
Shows the total number of samples and the total number of features for various cancer datasets.

Dataset	Number of samples	Number of features
Lung	1144	52
Squamous cell carcinoma	1144	52
Colon	62	2000
Leukaemia	72	7070
Breast	569	32

Table 2
Clinical feature distribution in Pan-Lung Cancer dataset samples.

Sample of Clinical features	Number of Samples
Cancer type: Lung Adenocarcinoma Lung Squamous Cell Carcinoma	660:484
Tumour stage: Stage I: Stage IV	8,246,321,. . .,8
Age range: <=60 years, >60 years	61:164
Gender: M, F	468:673
Smoking status	407:111
Smoking history	1100

Despite advancements in this field, additional research is required to better understand the process of metastasis and to improve the accuracy of diagnosis and prognosis for patients with lung cancer, since metastasis is the primary contributor to patient morbidity and mortality from cancer (Steeg, 2016). Therefore, identifying the characteristics that, whether clinical or genetic, most strongly influence the emergence of metastasis is essential for its early detection. Medical staff might then be able to choose the most effective treatments for LUAD patients with the help of this information. However, because there are few data sets with information on the metastatic stage of cancer patients and because these are variables of significant medical interest, there are many difficulties in studying this topic.

Additionally, it is important to consider a substantial disparity in the number of patients with and without metastasis, which poses a problem for precision medicine. Additional information is required than the kind of information currently available. In this paper, we provide a method to predict the presence of metastasis in Pan-Lung Cancer samples using a Deep Neural Learning (DNL) classifier model, with mutational load and pertinent clinical characteristics as predictors. Using this method, we could determine the ideal processing, training, and validation technique for DNL model while considering properties of this study's dataset. Benchmarking will be used to improve the correlations between genetic and clinical variables for cancer prognosis.

The first step in pre-processing is improving the features using Contrast Limit Adaptive Histogram Equalization. This was used as input for a prediction method using DNLC which has four covert layers in the network. Fig. 4 depicts the performance improvement during the training phase.

5.3. Colon cancer and leukaemia dataset

For this dataset, MLP has the highest classification accuracy for leukaemia and the lowest classification accuracy for colon cancer. Unlike with other datasets, DNLC more accurately classifies leukaemia and colon cancer. By running the simulation ten times with ten different beginning weight matrix values, we were able to achieve our experimental results.

The state-of-the-art CNN architecture (CNN based on SOAT), LeNet, for the detection of cancer is implemented by using the optimizer Adam. The veracity of the LeNet architectures' performance was validated and evaluated. This network for deep learn-

Table 3
Results of using 18 test samples to classify colon cancer dataset.

Classifier	Accuracy	F1 Score
CNN-based SOTA	0.90	0.903
RNN	0.81	0.83
DNLC	0.93	0.94

Using the Colon Cancer dataset, the DNLC approach was evaluated with an average accuracy of 93% and a F1 Score of 94%; compared to other CNN architectures based on state-of-the-art (SOTA) method, it had the highest efficiency.

Table 4
Results of using 22 test samples to classify leukaemia dataset.

Classifier	Accuracy	F1 Score
CNN-based SOTA	0.91	0.908
RNN	0.82	0.85
DNLC	0.94	0.96

Using the Leukaemia dataset, the DNLC approach was evaluated with an average accuracy of 94% and a F1 Score of 96%; compared to other CNN architectures based on state-of-the-art (SOTA) method, it had the highest efficiency.

Table 5
Results of using 12 test samples for the D1 Wisconsin Breast Cancer dataset.

Classifier	Accuracy	Sensitivity
URNN training	0.92.28	0.93.3
URNN testing	0.92.13	0.93.79
CNN-based SOTA	0.92.92	0.93.93
DNLC training	0.94.14	0.95.34
DNLC testing	0.95.45	0.95.87

Using the Wisconsin Breast Cancer dataset, the DNLC approach was evaluated with an average accuracy of 0.95.45% and a Sensitivity of 0.95.87%; compared to other CNN architectures based on state-of-the-art (SOTA) method, it had the highest efficiency.

Table 6
Results of using 12 test samples for the D2 Wisconsin Breast Cancer dataset.

Classifier	Accuracy	Sensitivity
URNN training	0.93.39	0.94.98
URNN testing	0.94.37	0.96.27
CNN-based SOTA	0.94.96	0.95.78
DNLC training	0.96.24	0.96.48
DNLC testing	0.95.97	0.95.88

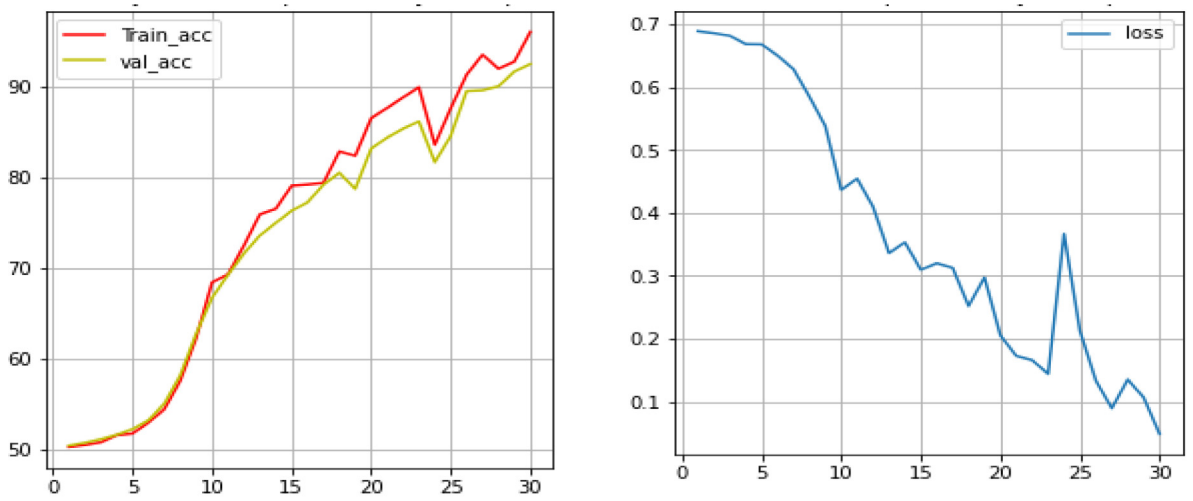

Fig. 4. The accuracy of DNLC for the different sample datasets.

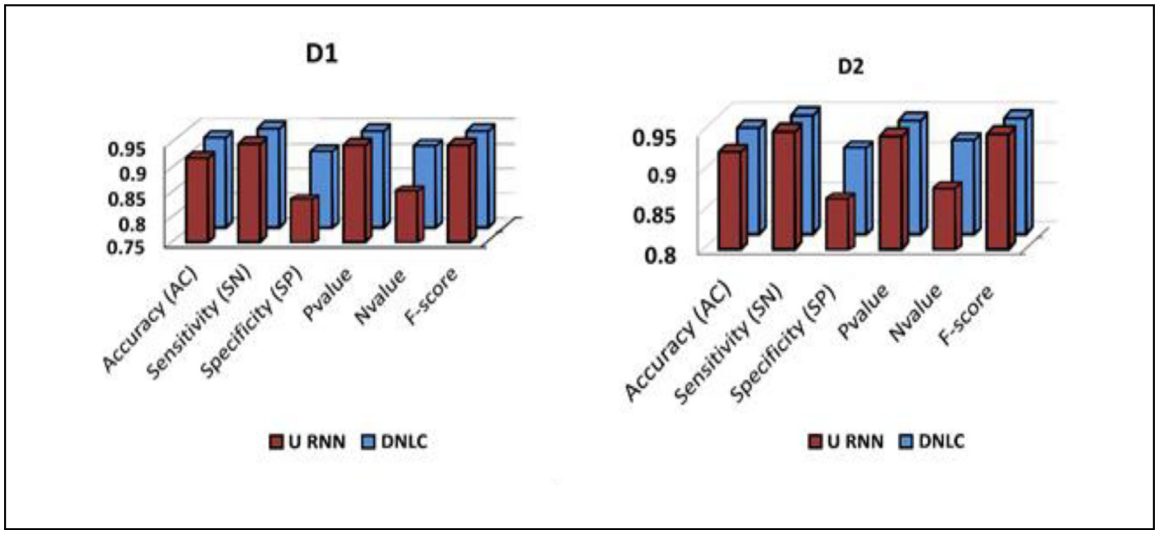

Fig. 5. Accuracy, sensitivity, specificity, precision, negative predictive, and F-score for URNN and DNLC.

ing employs binary cross-entropy loss. The result presented in this section was generated by the LeNet detection algorithm.

The result averages are presented in Tables 3 and 4. Information about the number of test samples is given in the title of each table, since the number varied for each dataset. It is immediately apparent that of the two cancer datasets, DNLC has the highest accuracy.

5.4. Breast cancer dataset

In the first experiment, we used a dataset from the UCI Machine Learning Repository (Wisconsin Breast Cancer) to differentiate between malignant (cancerous) and benign (non-cancerous) samples. Two labelled datasets were used in the trials (D1 Wisconsin Breast Cancer and D2 Wisconsin Breast Cancer).

The state-of-the-art CNN architecture (CNN based on SOAT), AlexNet for the detection of cancer is implemented by using the optimizer SGD. The veracity of the AlexNet architectures' performance was validated and evaluated. This network for deep learning employs binary cross-entropy loss. The result presented in this section was generated by the AlexNet detection algorithm.

This experiment aims to determine whether the proposed strategy should use unidirectional RNN (URNN) or bidirectional RNN (BRNN). DNLC is used to forecast the outcome. Tables 5 and 6, and Fig. 5 show the outcomes of using DNLC versus URNN. For all datasets used, these figures suggest that DNLC performs better than URNN.

Using the samples for the Wisconsin Breast Cancer dataset, the DNLC approach was evaluated with an average accuracy of 0.95.97 % and a Sensitivity of 0.95.88 %; compared to other CNN architectures based on state-of-the-art (SOTA) method, it had the highest efficiency.

The experimental results show that the suggested model performs better in terms of accuracy than earlier CNN and RNN models.

Using Breast Cancer dataset, their method was evaluated with an accuracy of 0.95.45% and a Sensitivity of 0.95.87%; compared to other SOTA methods, it had the highest efficiency.

6. Conclusions

Today, we understand that there are a variety of factors that have a substantial impact as causes of cancer. Because of this, it has become increasingly difficult to base a judgement on the categorization of a cancer on a single medical professional's opinion. Therefore, we require intelligent algorithms to provide support to medical professionals when diagnosing cancer. Additionally, several researchers have used predictive techniques like machine learning and deep learning to predict cancer prognoses. These techniques have been widely accepted for assessing the likelihood of patient survival. But there is widespread concern about the precision of such cancer prognoses. Deep neural learning (DNL) techniques are crucial for predicting a variety of diseases since they can swiftly predict outcomes from a sizable amount of clinical and genetic data. Our recommended method is built on deep neural learning. The following phases make up the Deep Neural Learning Cancer Prediction Model (DNLC). In the first stage, we employed Deep Network (DN) to choose the most appropriate set of characteristics from the datasets. The second stage involved training a deep neural network with samples of genomic or clinical data (DNN). The ability of the DNLC method to detect cancer at an early stage was assessed in the third stage. Five cancer datasets were used by DNLC for classification, including datasets for colon, lung adenocarcinoma, squamous cell carcinoma, breast, and leukaemia tumours. Our results show that the DNLC approach outperforms others in all situations, with an average accuracy of 93%.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abdeldayem, S. S., & Bourlai, T. (2018, December). ECG-based human authentication using high-level spectro-temporal signal features. In 2018 IEEE international conference on big data (big data) (pp. 4984-4993). IEEE.

Aly, G.H., Marey, M., El-Sayed, S.A., Tolba, M.F., 2021. YOLO based breast masses detection and classification in full-field digital mammograms. Comput. Methods Programs Biomed. 200, 105823.

Alzahrani, H., Venneri, A., 2015. Cognitive and neuroanatomical correlates of neuropsychiatric symptoms in Parkinson's disease: a systematic review. J. Neurol. Sci. 356 (1–2), 32–44.

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., & Zimmermann, T. "Software engineering for machine learning: A case study". In *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)* (pp. 291-300). IEEE.

Andersen, C.L., Jensen, J.L., Ørntoft, T.F., 2004. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance

estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. Cancer Res. 64 (15), 5245–5250.
Balloux, F., Brynildsrud, O.B., Van Dorp, L., Shaw, L.P., Chen, H., Harris, K.A., Eldholm, V., 2018. From theory to practice: translating whole-genome sequencing (WGS) into the clinic. Trends Microbiol. 26 (12), 1035–1048.
Buchman, D., Drozdov, M., Krilavičius, T., Maskeliūnas, R., Damaševičius, R., 2022. Pedestrian and animal recognition using doppler radar signature and deep learning. Sensors 22 (9), 3456.
Cheng, Z., Hu, X., Sun, Z., 2016. Microbial community distribution and dominant bacterial species analysis in the bio-electrochemical system treating low concentration cefuroxime. Chem. Eng. J. 303, 137–144.
Chong, Y., Wu, Y., Liu, J., Han, C., Gong, L., Liu, X., Li, S., 2021. Clinicopathological models for predicting lymph node metastasis in patients with early-stage lung adenocarcinoma: the application of machine learning algorithms. J. Thorac. Dis. 13 (7), 4033.
Cong, L., Feng, W., Yao, Z., Zhou, X., Xiao, W., 2020. Deep learning model as a new trend in computer-aided diagnosis of tumor pathology for lung cancer. J. Cancer 11 (12), 3615.
Dar, R.A., Rasool, M., Assad, A., 2022. Breast cancer detection using deep learning: datasets, methods, and challenges ahead. Comput. Biol. Med. 106073.
Deshmukh, P. B., & Kashyap, K. L. "Solution Approaches for Breast Cancer Classification through Medical Imaging Modalities Using Artificial Intelligence". In *Smart Trends in Computing and Communications*, Springer, Singapore. pp. 639-651, 2022
Dimitrakopoulos, C.M., Beerenwinkel, N., 2017. Computational approaches for the identification of cancer genes and pathways. Wiley Interdiscip. Rev. Syst. Biol. Med. 9 (1), e1364.
Ettayebi, K., Crawford, S.E., Murakami, K., Broughman, J.R., Karandikar, U., Tenge, V. R., Estes, M.K., 2016. Replication of human noroviruses in stem cell–derived human enteroids. Science 353 (6306), 1387–1393.
Glare, P., Virik, K., Jones, M., Hudson, M., Eychmuller, S., Simes, J., Christakis, N., 2003. A systematic review of physicians' survival predictions in terminally ill cancer patients. BMJ 327 (7408), 195.
Gonzalez-Ericsson, P. I., Stovgaard, E. S., Sua, L. F., Reisenbichler, E., Kos, Z., Carter, J. M., & International Immuno-Oncology Biomarker Working Group. "The path to a better biomarker: application of a risk management framework for the implementation of PD-L1 and TILs as immuno-oncology biomarkers in breast cancer clinical trials and daily practice". *The Journal of pathology*, *250*(5), 667-684, 2020
Hayes, D.F., Bast, R.C., Desch, C.E., Fritsche Jr, H., Kemeny, N.E., Jessup, J.M., Winn, R. J., 1996. Tumor marker utility grading system: a framework to evaluate clinical utility of tumor markers. JNCI: J. Nat. Cancer Inst. 88 (20), 1456–1466.
Helleday, T., Eshtad, S., Nik-Zainal, S., 2014. Mechanisms underlying mutational signatures in human cancers. Nat. Rev. Genet. 15 (9), 585–598.
Houssein, E.H., Emam, M.M., Ali, A.A., Suganthan, P.N., 2021. Deep and machine learning techniques for medical imaging-based breast cancer: a comprehensive review. Expert Syst. Appl. 167, 114161.
Huang, H.X., Li, J.C., Xiao, C.L., 2015. A proposed iteration optimization approach integrating backpropagation neural network with genetic algorithm. Expert Syst. Appl. 42 (1), 146–155.
Huang, X., Yu, R., Kang, J., He, Y., Zhang, Y., 2017. Exploring mobile edge computing for 5G-enabled software defined vehicular networks. IEEE Wirel. Commun. 24 (6), 55–63.
Kang, M., Ko, E., Mersha, T.B., 2022. A roadmap for multi-omics data integration using deep learning. Brief. Bioinform. 23 (1), bbab454.
Kanna, P.R., Santhi, P., 2022. Hybrid intrusion detection using mapreduce based black widow optimized convolutional long short-term memory neural networks. Expert Syst. Appl. 194, 116545.
Khachatryan, V., Sirunyan, A.M., Tumasyan, A., Adam, W., Asilar, E., Bergauer, T., Fonseca De Souza, S., 2016. "Event generator tunes obtained from underlying event and multiparton scattering measurements". Eur. Phys. J. C 76 (3), 1–52.
Khamparia, A., Bharati, S., Podder, P., Gupta, D., Khanna, A., Phung, T.K., Thanh, D.N. H., 2021. Diagnosis of breast cancer based on modern mammography using hybrid transfer learning. Multidimens. Syst. Signal Processing 32, 747–765.
Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Getz, G., 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 499 (7457), 214–218.
Laxminarayanamma, K., Krishnaiah, R.V., Sammulal, P., 2022. Enhanced CNN model for pancreatic ductal adenocarcinoma classification based on proteomic data. Ingénierie des Systèmes d'Information 27 (1).
Loyola-González, O., García-Borroto, M., Medina-Pérez, M.A., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A., Ita, G.D., 2013. In: June). An empirical study of oversampling and undersampling methods for LCMine an emerging pattern based classifier. Springer, Berlin, Heidelberg, pp. 264–273.
Madabhushi, A., Agner, S., Basavanhally, A., Doyle, S., Lee, G., 2011. Computer-aided prognosis: predicting patient and disease outcome via quantitative fusion of multi-scale, multi-modal data. Comput. Med. Imaging Graph. 35 (7–8), 506–514.
Mafarja, M.M., Mirjalili, S., 2017. Hybrid whale optimization algorithm with simulated annealing for feature selection. Neurocomputing 260, 302–312.
Mela, K., Tiainen, T., Heinisuo, M., 2012. Comparative study of multiple criteria decision making methods for building design. Adv. Eng. Inf. 26 (4), 716–726.
Meuten, D.J., Moore, F.M., Donovan, T.A., Bertram, C.A., Klopfleisch, R., Foster, R.A., Whitley, D., 2021. International guidelines for veterinary tumor pathology: a call to action. Vet. Pathol. 58 (5), 766–794.
Mohammed, M., Mwambi, H., Mboya, I.B., Elbashir, M.K., Omolo, B., 2021. A stacking ensemble deep learning approach to cancer type classification based on TCGA data. Sci. Rep. 11 (1), 1–22.
Mohd Amiruddin, A.A.A., Zabiri, H., Taqvi, S.A.A., Tufa, L.D., 2020. Neural network applications in fault diagnosis and detection: an overview of implementations in engineering-related systems. Neural Comput. & Applic. 32, 447–472.
Murtaza, G., Shuib, L., Abdul Wahab, A.W., Mujtaba, G., Nweke, H.F., Al-garadi, M.A., Azmi, N.A., 2020. Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges. Artif. Intell. Rev. 53 (3), 1655–1720.
Naik, A., Edla, D.R., 2021. Lung nodule classification on computed tomography images using deep learning. Wirel. Pers. Commun. 116, 655–690.
Nhu, V.H., Shirzadi, A., Shahabi, H., Singh, S.K., Al-Ansari, N., Clague, J.J., Ahmad, B.B., 2020. "Shallow landslide susceptibility mapping: a comparison between logistic model tree, logistic regression, naïve bayes tree", artificial neural network, and support vector machine algorithms. Int. J. Environ. Res. Public Health 17 (8), 2749.
Özkaya, U., Seyfi, L., Öztürk, Ş., 2021. Dimension optimization of multi-band microstrip antennas using deep learning methods. Pamukkale Üniversitesi Mühendislik Bilimleri Dergisi 27 (2), 229–233.
Öztürk, Ş., Çukur, T., 2022. Deep clustering via center-oriented margin free-triplet loss for skin lesion detection in highly imbalanced datasets. IEEE J. Biomed. Health Inform. 26 (9), 4679–4690.
Öztürk, Ş., Özkaya, U., Akdemir, B., & Seyfi, L. (2018, November). Weighting and Classification of Image Features using Optimization Algorithms. In 2018 International Symposium on Fundamentals of Electrical Engineering (ISFEE) (pp. 1-6). IEEE
Ozturk, S., Unal, Y., 2020. A two-stage whale optimization method for classification of parkinson's disease voice recordings. Int. J. Intell. Syst. Appl. Eng. 8 (2), 84–93.
Pao, W., Ladanyi, M., 2007. Epidermal growth factor receptor mutation testing in lung cancer: searching for the ideal method. Clin. Cancer Res. 13 (17), 4954–4955.
Pohekar, S.D., Ramachandran, M., 2004. Application of multi-criteria decision making to sustainable energy planning—a review. Renew. Sustain. Energy Rev. 8 (4), 365–381.
Polat, K., Güneş, S., 2007. Automatic determination of diseases related to lymph system from lymphography data using principles component analysis (PCA), fuzzy weighting pre-processing and ANFIS. Expert Syst. Appl. 33 (3), 636–641.
Qian, Y., Zhang, Y., Ma, X., Yu, H., Peng, L., 2019. EARS: emotion-aware recommender system based on hybrid information fusion. Information Fusion 46, 141–146.
Riazalhosseini, Y., Lathrop, M., 2016. Precision medicine from the renal cancer genome. Nat. Rev. Nephrol. 12 (11), 655–666.
Sampedro, C., Martinez, C., Chauhan, A., Campoy, P., 2014. In: July). A supervised approach to electric tower detection and classification for power line inspection. IEEE, pp. 1970–1977.
Selvanambi, R., Natarajan, J., Karuppiah, M., Islam, S.H., Hassan, M.M., Fortino, G., 2020. Lung cancer prediction using higher-order recurrent neural network based on glowworm swarm optimization. Neural Comput. & Applic. 32, 4373–4386.
Senthil, S., Ayshwarya, B., 2018. Lung cancer prediction using feed forward back propagation neural networks with optimal features. Int. J. Appl. Eng. Res. 13 (1), 318–325.
Shaffie, A., Soliman, A., Fraiwan, L., Ghazal, M., Taher, F., Dunlap, N., El-Baz, A., 2018. A generalized deep learning-based diagnostic system for early diagnosis of various types of pulmonary nodules. Technol. Cancer Res. Treat. 17.
Sharan, E. S., Kumar, K. S., & Madhuri, G. "Conceal Face Mask Recognition Using Convolutional Neural Networks". In *2021 6th International Conference on Communication and Electronics Systems (ICCES)* (pp. 1787-1793) IEEE.
Sharma, M., Singh, S.K., Agrawal, P., Madaan, V., 2016. Classification of clinical dataset of cervical cancer using KNN. Indian J. Sci. Technol. 9 (28), 1–5.
Shukla, A.K., Singh, P., Vardhan, M., 2019. A hybrid framework for optimal feature subset selection. J. Intell. Fuzzy Syst. 36 (3), 2247–2259.
Statnikov, A., Aliferis, C.F., Tsamardinos, I., Hardin, D., Levy, S., 2005. A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. Bioinformatics 21 (5), 631–643.
Steeg, P.S., 2016. Targeting metastasis. Nature reviews cancer 16 (4), 201–218.
Travis, W.D., Brambilla, E., Riely, G.J., 2013. New pathologic classification of lung cancer: relevance for clinical practice and clinical trials. J. Clin. Oncol. 31 (8), 992–1001.
Tubishat, M., Ja'afar, S., Alswaitti, M., Mirjalili, S., Idris, N., Ismail, M.A., Omar, M.S., 2021. Dynamic salp swarm algorithm for feature selection. Expert Syst. Appl. 164, 113873.
Valdez, F., Melin, P., & Castillo, O. "Evolutionary method combining particle swarm optimization and genetic algorithms using fuzzy logic for decision making". In *2009 IEEE International Conference on Fuzzy Systems* (pp. 2114-2119). IEEE.
van Vliet, M.H., Reyal, F., Horlings, H.M., van de Vijver, M.J., Reinders, M.J., Wessels, L.F., 2008. Pooling breast cancer datasets has a synergetic effect on classification performance and improves signature stability. BMC Genomics 9 (1), 1–22.
Withnell, E., Zhang, X., Sun, K., Guo, Y., 2021. XOmiVAE: an interpretable deep learning model for cancer classification using high-dimensional omics data. Brief. Bioinform. 22 (6), bbab315.
Wu, F., Souza, A., Zhang, T., Fifty, C., Yu, T., & Weinberger, K. "Simplifying graph convolutional networks". In *International conference on machine learning* (pp. 6861-6871), (2019, May). PMLR.

Xu, J., & Zhang, J. "Exploration-exploitation trade-offs in metaheuristics: Survey and analysis". In *Proceedings of the 33rd Chinese control conference* (pp. 8633-8638), (2014, July). IEEE.

Yang, W.S., SriRamaratnam, R., Welsch, M.E., Shimada, K., Skouta, R., Viswanathan, V.S., Stockwell, B.R., 2014. Regulation of ferroptotic cancer cell death by GPX4. Cell 156 (1–2), 317–331.

Zamry, N.M., Zainal, A., Rassam, M.A., Alkhammash, E.H., Ghaleb, F.A., Saeed, F., 2021. Lightweight anomaly detection scheme using incremental principal component analysis and support vector machine. Sensors 21 (23), 8017.

Zováthi, B.H., Mohácsi, R., Szász, A.M., Cserey, G., 2022. Breast tumor tissue segmentation with area-based annotation using convolutional neural network. Diagnostics 12 (9), 2161.

Haitham Elwahsh received his B.Sc. degree in computer science, from Faculty of Science-Qena, South Valley University in 2004 and his M.Sc. degree in Computer Science in 2012 from Menoufia University. He received his Ph.D. degree of Science in Mathematics and Computer Science, Port Said University, Egypt in 2019. He is currently an assistant professor at Computer Science Department, Faculty of Computers and Information, Kafrelsheikh University, Kafrelsheikh 33516, Egypt. He is a reviewer in highly ranked journal such as IEEE Access Journal, Peerj Computer Science, Computer Systems Science and Engineering, Journal of Ambient Intelligence and Humanized Computing. His research interests include Network Security, Cyber Security, Mobile Ad-hoc Networks (MANETs), Machine Learning, Artificial Intelligence and image processing.

Medhat A. Tawfeek received the B.Sc., M.Sc. and PhD in Computer Science from Menoufia University, Faculty of Computers and Information in 2005, 2010 and 2015 respectively. He is presently an assistant professor in the department of Computer Science, Faculty of Computers and Information, Menoufia University, Egypt. He also holds the same position in in the department of Computer Science, College of Computer and Information Sciences, Jouf University, KSA. His research interest includes cloud computing, internet of things, smart card security, machine learning, distributed system, and fault tolerance.

A. A. Abd El-Aziz has completed Ph.D. degree in June 2014 in Information Science Technology from Anna University, Chennai-25, India. He has received B.Sc., and Master of Computer Science in 1999 and 2006 respectively from Faculty of Science, Cairo University. Now, he is an Ass. Prof in the FGSSR, Cairo University, Egypt. He has 21 years' experience in teaching at Cairo University, Egypt. His research interests are database system, database security, Cloud computing, Big data, ML and XML security. He has authored/coauthored over 60 research publications in peer-reviewed reputed journals and conference proceedings. He advised more than 10 master's graduates. Moreover, he has also served as a technical program committee member for many workshops and conferences and he has served as a reviewer for various international journals.

Mahmood A. Mahmood, completed Ph.D. degree in June 2014 in Computer Science and Information from Cairo University, Faculty of Graduate Studies for Statistical Research (FGSSR). I have received my B.Sc., 1998 from Faculty of Science, Cairo University and Master of Computer Science and Information in 2009 from Statistical Studies and Research, Cairo University. Now, I am an Associate Professor in the FGSSR, Cairo University, Egypt. I have 18 years, experience in Teaching at Cairo University, Egypt. Currently, I'm working as an assistant professor at Jouf University, Saudi Arabia. My research interests include Artificial Intelligent, Data Mining, Recommender Systems and Multimedia.

Maazen Alsabaan received his B.S. degree in electrical engineering, from King Saud University, Saudi Arabia, in 2004, and his M.A.Sc. and Ph.D. degrees in electrical and computer engineering from the University of Waterloo, Canada, in 2007 and 2013, respectively. He is currently an assistant professor in the Department of Computer Engineering, King Saud University (KSU), Saudi Arabia. He was the Chairman of the department from 2015 to 2018. Dr. Alsabaan serves as a consultant for different agencies and has been awarded many grants from KSU and King Abdulaziz City for Science and Technology (KACST). He is an associate editor of Journal of Circuits, Systems, and Computers. His current research interests include wireless communications and networking, surveillance systems, vehicular networks, green communications, intelligent transportation systems, and cybersecurity.

Engy A. El-shafeiy is the Head of electronic exams, University of Sadat City. She received a Ph.D. degree in 2019 in Big data mining and AI from computers engineering department -Faculty of engineering, Mansoura University, a M. SC degree in distribution Control Systems Engineering in 2008 from computers engineering department -Faculty of engineering, Mansoura University, and received B.Sc. in Computers Engineering and Systems in 2001 from Faculty of Computers Engineering Department, Faculty of engineering, Mansoura University in Egypt. She is currently an assistant professor at Department of Computer Science, Faculty of Computers and Artificial Intelligence, University of Sadat City, Sadat City, Egypt. A Reviewer for different highly ranked International Journals such as Expert Systems with Applications (Elsevier), journal of big data in springer. Her research focuses on Artificial Intelligence techniques, Hybridization of different machine learning techniques, Big Data analysis, Artificial Neural Networks, Optimization algorithms, Wireless Networks, network security, intelligent environment applications, Biomedical and software engineering.