Received April 8, 2020, accepted April 29, 2020, date of publication May 6, 2020, date of current version May 27, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2992907

Deep Learning for Multi-Tissue Cancer Classification of Gene Expressions (GeneXNet)

TAREK KHORSHED, (Member, IEEE), MOHAMED N. MOUSTAFA, (Member, IEEE), AND AHMED RAFEA

Department of Computer Science and Engineering, The American University in Cairo, New Cairo 11835, Egypt

Corresponding author: Tarek Khorshed (tarek_khorshed@aucegypt.edu)

ABSTRACT Cancer classification using gene expressions is extremely challenging given the complexity and high dimensionality of the data. Current classification methods typically rely on samples collected from a single tissue type and perform a prerequisite of gene feature selection to avoid processing the full set of genes. These methods fall short in taking advantage of genome-wide next generation sequencing technologies which provide a snapshot of the whole transcriptome rather than a predetermined subset of genes. We propose a deep learning framework for cancer diagnosis by developing a multi-tissue cancer classifier based on whole-transcriptome gene expressions collected from multiple tumor types covering multiple organ sites. We introduce a new Convolutional Neural Network architecture called Gene eXpression Network (GeneXNet), which is specifically designed to address the complex nature of gene expressions. Our proposed GeneXNet provides capabilities of detecting genetic alterations driving cancer progression by learning genomic signatures across multiple tissue types without requiring the prerequisite of gene feature selection. Our model achieves 98.9% classification accuracy on human samples representing 33 different cancer tumor types across 26 organ sites. We demonstrate how our model can be used for transfer learning to build classifiers for tumors lacking sufficient samples to be trained independently. We introduce visualization procedures to provide biological insight on how our model is performing classification across multiple tumors.

INDEX TERMS Cancer classification, convolutional neural networks, deep learning, gene expressions, next generation sequencing, RNA sequencing, transfer learning.

I. INTRODUCTION

The World Health Organization reports that cancer is a leading cause of death worldwide accounting for an estimated 9.6 million deaths in 2018 [1]. Despite this dramatic impact, between 30-50% of cancer death cases can be prevented through early detection and treatment [2]. Advancements in cancer classification and prediction play an important role in early detection since a major challenge in cancer treatment is that patients are diagnosed at very late stages where appropriate interventions become less effective and full curative treatment is no longer achievable [3].

Gene expressions have been extensively used in cancer classification [6]–[13]. Technological advances in structural genomics have allowed studying the full set of DNAs in the human genome [3], [21]. Next generation sequencing (NGS) methods such as whole-genome DNA sequencing and Total RNA sequencing are considered revolutionary technologies for studying genetic changes in cancer [18], [23]. These technologies provide great potential for cancer classification and better understanding of tumor progression given their ability to sequence thousands of genes at one time and detect multiple types of genomic alterations [16], [17], [21]. They provide capabilities for comparing the sequence of DNA and RNA in cancer cells with that in normal cells to identify genetic changes that may be driving the growth of a tumor. Gene expression analysis using Total RNA sequencing provides a snapshot of the whole transcriptome rather than a predetermined subset of genes, enables testing multiple genes simultaneously and can detect both coding plus multiple forms of noncoding RNA [18]. These methods have eliminated many limitations involved in microarray based experiments that were traditionally used for gene expression analysis [18], [21], [23].

Despite all these potential capabilities, cancer classification using gene expressions produced from Total RNA sequencing is extremely challenging given the complexity and massive amount of genetic data that is

The associate editor coordinating the review of this manuscript and approving it for publication was Gustavo Callico.

produced [16], [17], [21], [22], [30]. The magnitude of data variants obtained from RNA Sequencing is exponential which makes it difficult for traditional machine learning approaches to evaluate genetic variants for disease prediction [3], [18], [19]. Gene expression data is characterized by being very high in dimensionality in terms of having a very large number of features representing the genes, and a very small number of training data representing the patient samples [7], [27]. Complexity is also due to the fact that only a small subset of genes might be driving the cancer tumor progression [1], [3].

Current cancer classification methods avoid processing the full set of genes to overcome these complexities and are mainly based on performing a process of gene feature selection as a prerequisite to the classifier learning process [24]–[27]. Gene feature selection will allow the learning process to proceed, but the resulting classifier will not have the opportunity to learn the molecular signatures of genes which have been excluded and their influence on the underlying cancer tumor [28], [29]. Current classification methods based on gene feature selection are not optimal for early cancer diagnosis. This is because these methods fall short in taking the full advantage of DNA and RNA sequencing technologies to discover the correlated patterns between genes across the full set of DNAs in the human genome and to detect multiple types of genetic alterations that may be driving the growth of a tumor across the whole transcriptome rather than a predetermined subset of genes [4]. Another limitation of current methods is that they typically rely on gene expressions collected from a single cancer tissue type based on the same anatomical site of origin. This approach does not utilize the full potential of emerging whole-genome sequencing technologies and data produced by large-scale genomic projects that produce detailed molecular characterizations of thousands of tumors using genome-wide platforms [30]. Recent studies which have performed an integrated multi-platform analysis across multiple cancer types have revealed molecular classification within and across tissues of origin [4], [5]. The results of these studies have recommended that the traditional approach of anatomic cancer classification should be supplemented by classification based on molecular alterations shared by tumors across different tissue types [4].

This has motivated our research for early diagnosis of cancer by leveraging the latest deep learning methods to develop a comprehensive multi-tissue cancer classifier. Our proposed classifier is based on molecular signatures of whole-transcriptome wide gene expressions, that are collected from human samples representing multiple cancer tissue types covering multiple organ sites of origin. Our approach using deep learning eliminates the need for discovering a predefined subset of genes by combining the process of gene feature selection and classification into one end-to-end learning system. We propose a new Convolutional Neural Network architecture called "Gene eXpression Network" (GeneXNet) which is specifically designed to learn the complex nature of whole-transcriptome gene expressions and which gives the opportunity to design cancer classifiers with capabilities of detecting more complex types of genetic alterations by learning the genomic signatures shared across multiple cancer tissue types. To our knowledge, this is the first effort to develop a multi-tissue cancer classifier based on a full set of whole-transcriptome wide gene expressions collected from tumors across different tissue types without requiring a prerequisite process of gene feature selection. We demonstrate how our model can perform transfer learning to build classifiers for other types of cancer tumors which are lacking sufficient patient samples to be trained independently. We introduce visualization procedures to provide more biological insight on how our model is performing cancer classification across multiple tumor types. We visualize gene localization maps highlighting the important regions in the gene expressions influencing the tumor class prediction. We also visualize the molecular clusters formed by intermediate gene expression feature maps learned by the network which helps in revealing the genomic relationships of gene expressions that are influential in the tumor progression.

II. RELATED WORK

Gene expressions have been extensively used in cancer classification [6]–[13]. Transcription produces what is referred to as precursor messenger RNA (pre-mRNA) which undergoes further modifications leading to mature mRNA [1]. By collecting mRNA samples for tumors of known classes, supervised learning can be used to build discriminative models which can learn the gene patterns of the underlying disease and then be used to predict the tumor class of new patient samples which were not previously diagnosed [1]. This is considered a great achievement as there are many microarray experiments which demonstrate how it was possible to distinguish between certain cancer types using data classification even though they are clinically indistinguishable [1], [56], [57].

Current methods for cancer classification follow the approach of feature engineering and are based on applying innovative gene feature selection techniques as a prerequisite to the classifier learning process to discover a small subset of informative genes which are discriminative among the tumor being analysed [24], [26]. Gene selection methods can be generally classified into filtering, wrapping and embedded methods [27]. The accuracy of such a classifier depends heavily on the successful identification of these discriminative features [28], [29].

Substantial work has been done for cancer classification by performing gene feature selection and building on traditional machine learning methods such as Support Vector Machines [13], [15], [25], Random Forests [12], Decision Trees [14], AdaBoost [9], K-Nearest Neighbor [12] and Genetic algorithms [7], [9]. Many other techniques which combine gene feature selection and cancer classification have also been proposed for gene expressions in addition to other types of Omics data [6], [7], [8], [10], [11], [58].

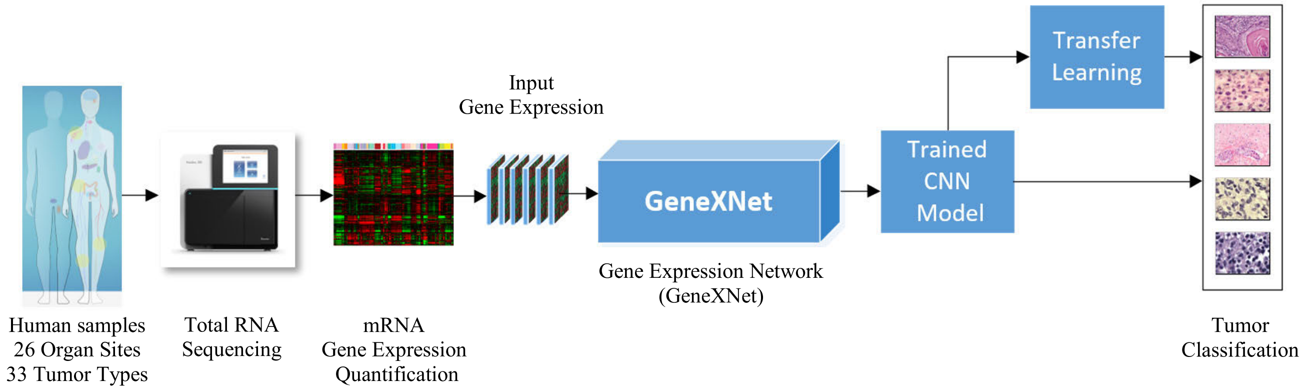

FIGURE 1. Deep learning system architecture. The system starts with data collection of cancer tumors using Total RNA Sequencing, followed by training our proposed CNN then performing tumor classification.

III. PROPOSED APPROACH

A. DEEP LEARNING FOR MULTI-TISSUE CANCER CLASSIFICATION

Current methods for cancer classification are based on gene feature selection as a prerequisite to the classifier learning process. Our approach using Deep Learning provides an alternative solution to feature engineering and eliminates the need for discovering a predefined subset of genes. This is achieved by combining the process of gene feature selection and classification into one end-to-end learning system using the whole set of transcriptome wide gene expressions collected from tumors across different tissue types. Our proposed Convolutional Neural Network (CNN) architecture combines multiple layers of non-linear building blocks which transform the gene expression data into a representation at a higher more abstract level. This allows the network to automatically learn the molecular patterns of expressed genes which are influencing the tumors and use that to amplify the discrimination score for classification. The advantage is that the classifier will not be limited to learning the molecular characterization of a single tissue type but will have the capability of detecting more complex types of genomic alterations by learning the genetic signatures collected from multiple tumors and across multiple cancer tissue types. Another major advantage of our approach is that it allows performing very efficient transfer learning by reusing the molecular signatures learned by the trained networks.

B. DEEP LEARNING SYSTEM ARCHITECTURE

A schematic diagram of our end-to-end learning system architecture is shown in Fig. 1. The first section represents the data collection and preparation process. It depends on collecting human samples representing multiple types of cancer tumors collected from multiple tissues spanning different organs across the body. The next step performs gene expression quantification using a Next Generation Sequencing procedure. Total RNA sequencing is performed for measuring gene expression quantification across the whole-transcriptome and extracting coding mRNA. The gene expression data is normalized and converted into a representation suitable for feeding it as input data to our deep learning model.

The second section of our learning system represents building and training a deep Convolutional Neural Network to automatically learn the molecular signatures of the full set of whole-transcriptome gene expressions and produce a trained model which can be used for classification of cancer tumors. Our model, which we refer to as ‘‘Gene eXpression Network’’ (GeneXNet), relies on building an architecture with multiple layers of non-linear functions which transform the gene expression data into feature maps to increase the level of accuracy and invariance of the selected gene features [53]. The genetic signatures learned by the feature maps in the deep layers, eliminate the need for the traditional prerequisite process of gene feature selection because they are insensitive to any insignificant genes or irrelevant variations in the gene expressions [47], [53]. We train the model using supervised learning by feeding the collected human samples as input and producing an output probability score for each labelled category of cancer tumors. We define a cross-entropy loss function suitable for gene expression data that measures the error between the network input and the desired output, then we use stochastic gradient descent optimization and back-propagation [48] to adjust the network weights.

C. CONVOLUTIONAL NEURAL NETWORKS (CNNs)

Convolutional Neural Networks (CNNs) have contributed to many record breaking achievements especially in the areas of computer vision and image recognition [31]–[36]. The development of new CNN architectures to improve accuracy and performance continues to be an active research area such as AlexNet [52], VGGNet [50], GoogLeNet [49], InceptionNet [45], ResNet [36], [46], DenseNet [33], MobileNet [31], [35], SENet [34] and NasNet [32]. CNNs are made of multiple layers each arranged in a 3D volume of neurons where each layer transforms the volume using a non-linear transformation. CNNs differ from fully connected networks in that neurons are only connected to a small region

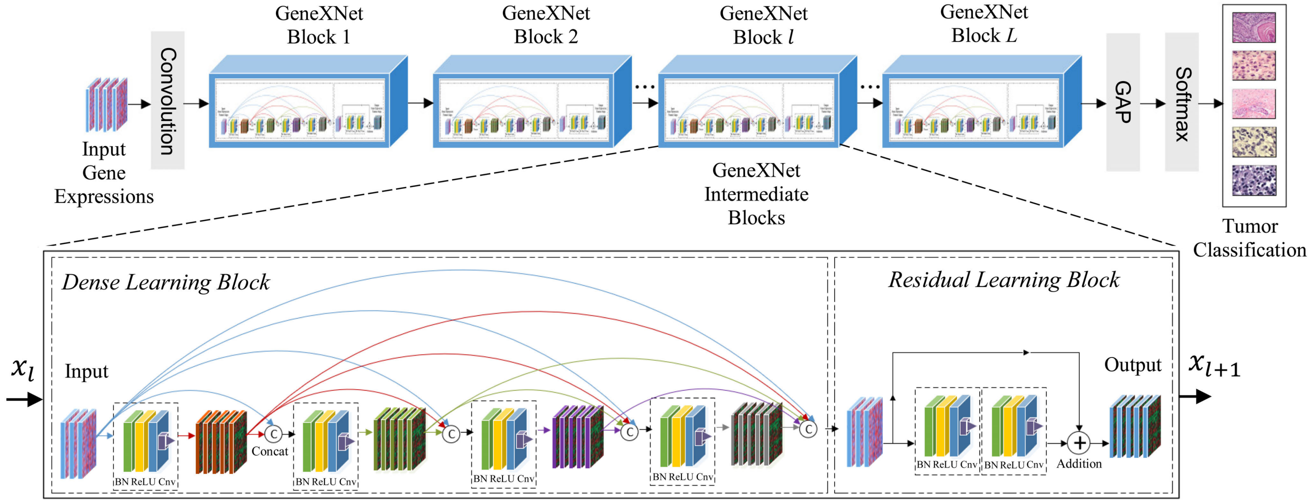

FIGURE 2. Gene eXpression Network (GeneXNet) architecture. Our proposed CNN incorporates multiple layers of building blocks which include a combination of Dense and Residual learning sub-blocks.

in the previous layer. The convolution operation performs a dot product between a sliding filter and the input across the full depth of the volume to produce an activation map [47].

The motivation behind using CNNs for classification of cancer tumors using gene expressions is that the convolution operation is very suitable for the high dimensional and sparse nature of the data. Since the input data has a very high dimensionality, it is not practical to use traditional kernel learning methods and fully connected networks since the resulting models will have a huge number of parameters to be learned which makes the learning process infeasible [47].

D. GENE EXPRESSION REPRESENTATION FOR CNNs

To train our CNN model using the cancer tumor samples, we first need to represent the gene expressions in a format suitable for the network input. Given N tumor samples each having G features representing the full set of genes produced by the whole-transcriptome sequencing procedure, we can represent the gene expressions in an equivalent 2D matrix of real numbers with dimensions (N, G) which stores the normalized gene expressions such that the value in cell X_{ij} represents the expression level measured for gene (j) in the patient sample (i). We convert each sample into the equivalent 3D volume of genes with dimensions (Width, Height, Depth) to make it suitable as input to our CNN model. The training data for the N samples can then be represented by the 4D input matrix of real numbers with dimensions (N, W, H, D).

E. GENE EXPRESSION NETWORK (GeneXNet) ARCHITECTURE

In this section we describe the detailed architecture of our proposed CNN model. Recent benchmark results obtained by deep CNNs for image recognition tasks have demonstrated that network depth is of great importance for feature extraction and have managed to achieve outstanding results by designing networks with deeper and more complex architectures [36], [46]. These models were able to exploit deep architectures because of the availability of large training datasets such as ImageNet which contains over 1 million training images [52]. Training deep models requires large amounts of training data to avoid common problems such as overfitting, vanishing gradients and degradation of accuracy [36], [46]. Applying the same deep CNN architectures for classification of gene expression data is not an evident task since it faces two conflicting problems. On one hand, we need to benefit from deep network architectures to efficiently extract the molecular signatures of the large number of genes so that our classifier can accurately generalize when presented with tumor data from multiple tissue types. But on the other hand, the lack of sufficient human training samples, which could be in the range of only a few hundred samples, implies great challenges for training deep networks and results in overfitting during training which implies using smaller more compact networks. We attempted to build an end-to-end learning system for classification without performing the prerequisite process of gene feature selection by using some of the available state-of-the-art CNN models which have been specifically designed for computer vision tasks. Our experimental results have shown that training these deep models suffered from overfitting when presented with the underlying dataset that includes the full set of transcriptome gene expressions collected from tumors across different tissue types. The dataset did not have sufficient training samples to train these deep models and achieve the required accuracy.

To solve these conflicting problems, we propose a new CNN architecture which we refer to as Gene eXpression Network (GeneXNet) shown in Fig. 2. Our network is designed to specifically address the complex nature of gene expressions and addresses the lack of training samples by incorporating multiple layers of building blocks which we refer

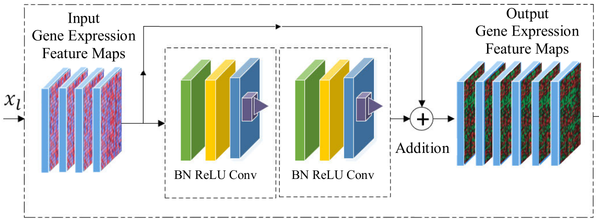

FIGURE 3. **Residual learning block of Gene eXpression Network.**

to as GeneXNet blocks. These blocks are motivated from both deep residual learning networks [36], [46] and densely connected convolutional networks [33] and are formed by merging together two different types of learning sub-blocks.

To formulate our building block, we define our network to have L layers of blocks where the non-linear transformation of gene expressions can be denoted by G_l and defined as:

$$x_{l+1} = G_l(x_l, W_l) \tag{1}$$

$$W_l = \left\{ w_{l,i} | 1 \leq i \leq K_l \right\} \tag{2}$$

where l is the index of the block, W_l represents the set of weights and biases of the l^{th} block, $w_{l,i}$ represents the weights of the i^{th} convolutional layer in the l^{th} block, K_l represents the number of convolution layers in the l^{th} block and x_l, x_{l+1} represent the input and output features of the l^{th} block. We apply "pre-activation" of weight layers as in [36] by defining the transformation at each layer as a sequence of multiple operations which are Batch Normalization (BN) [44], Rectified Linear Unit (ReLU) [38] and Convolution.

If gene expression data flows through the network using only the transformation in (1), that would be following the traditional approach for CNN layers. Deep residual learning provides a framework for more efficiently training deep networks by reformulating the layers as residual learning functions with reference to the layer inputs [46]. Empirical results have shown that residual learning helps to avoid degradation in performance accuracy as the depth of the network increases [46]. Residual networks have achieved excellent performance in many image recognition and object detection tasks, where networks with over 150 layers have been trained on ImageNet [52] and managed to achieve substantial accuracy gains in comparison to normal networks which simply stack consecutive layers [36]. To make use of residual learning we reformulate our building block by implementing the non-linear transformation of gene expressions G_l as a residual function defined as:

$$x_{l+1} = f_l\left[G_l(x_l, W_l) + M(x_l)\right] \tag{3}$$

where G_l is a residual function for the l^{th} block, $M(x_l)$ is a mapping which bypasses the non-linear transformation and f_l represents a mapping function of the input and output features of the l^{th} block. The simplest form of residual learning can be realized by choosing f_l to be a Rectified Linear Unit (ReLU) [38] and also introducing identity skip connections which are equivalent to choosing $M(x_l)$ as an identity mapping so that $M(x_l) = x_l$. Another formulation can be realized by implementing both $M(x_l)$ and f_l as identity mappings. We apply the later formulation which has shown to improve accuracy by creating a more direct path for information propagation and allowing the signal to propagate more directly from one unit to any other unit in the forward and backward passes [36]. The resulting non-linear transformation of gene expressions and the gradient of the loss function can then be expressed recursively as:

$$x_L = x_l + \sum_{i=l}^{L-1} G_i(x_i, W_i) \tag{4}$$

$$\frac{\partial \varepsilon}{\partial x_l} = \frac{\partial \varepsilon}{\partial x_L}\frac{\partial x_L}{\partial x_l} = \frac{\partial \varepsilon}{\partial x_L}\left[1 + \frac{\partial}{\partial x_l}\sum_{i=l}^{L-1} G_i(x_i, W_i)\right] \tag{5}$$

where x_L represents the output features of the network with L layers of blocks, ε is the loss function and $\partial\varepsilon/\partial x_l$ is the gradient obtained by applying the chain rule and backpropagation [36]. The residual function G_i is implemented as in (1) by applying two or more weight layers each using pre-activation and the sequence of multiple operations BN, ReLU and convolution. The resulting sub-block is shown in Fig. 3 which we refer to as the *Residual Learning block*. We also experiment with applying a bottleneck architecture [36], [46], by modifying the design of this block to have three layers instead of two in the form of (1 × 1), (3 × 3) and (1 × 1) convolutions. Since we are using the full set of whole-transcriptome genes, the role of the (1 × 1) convolution is to enhance computational efficiency by reducing the large dimensions of the intermediate feature maps before applying the (3 × 3) convolution and then restore them back again [36].

Despite the strong advantages of residual learning networks in allowing the gradient to flow directly through the skip connections, there have been other proposed approaches to use stochastic depth to improve the training of deep residual networks by dropping layers randomly during training [37]. This has led to different intuitions that there might be a great amount of redundancy in deep residual networks and that not all the layers are required [33]. Densely connected convolutional networks (DenseNets) [33] exploit the potential of the network through feature reuse as an alternative to deep or wide architectures by connecting all layers with matching feature-map sizes directly with each other. This design consideration is very important for our task, since one of the biggest challenges in our work is to build a multi-tissue cancer classifier that can benefit from deep network architectures to efficiently extract the molecular signatures of large number of genes, without facing severe overfitting or degradation in performance due to the lack of sufficient human training samples. This has inspired us to further reformulate the design of our GeneXNet building block and augment its learning capability by introducing additional dense layers that precede the residual learning layers. The dense layers follow a similar approach as in DenseNets [33].

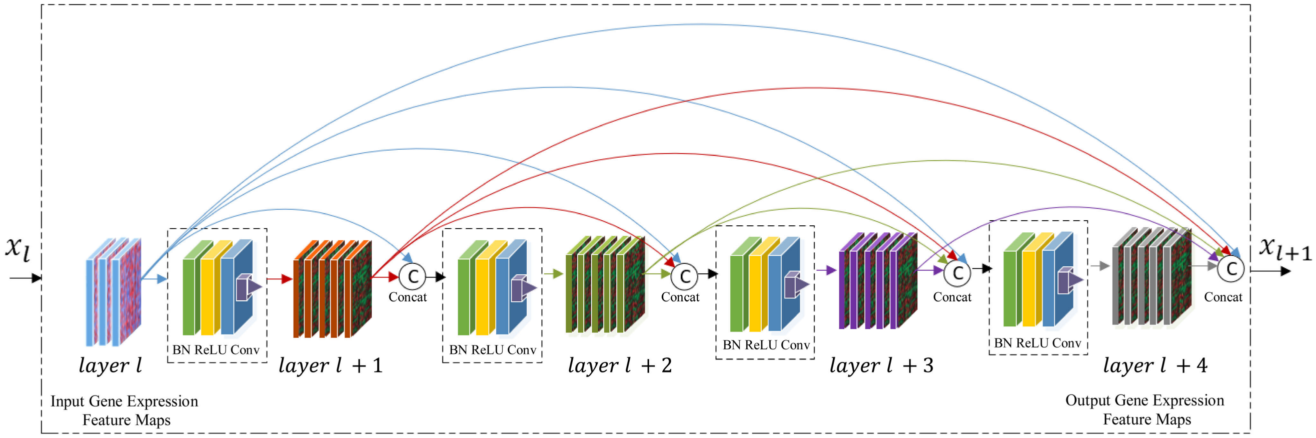

FIGURE 4. Dense learning block of Gene eXpression Network.

The design of our dense layers is implemented by passing additional inputs into each layer from all preceding layers and passing the feature maps of each layer to all subsequent layers. Our aim from this design is to provide each layer with direct access to the gradients from the loss functions and the original input signal which can potentially improve flow of information throughout the network. Our additional dense layers are formulated as:

$$x_{l+1} = G_l\left(Concat\left[x_1, x_2, x_3, \ldots, x_l\right]\right) \quad (6)$$

where x_{l+1} represents the output of the l^{th} block, $Concat\left[x_1, \ldots, x_l\right]$ represents the concatenation of the gene expression feature maps resulting from all preceding layers and G_l represents the same transformation as in (1) which applies pre-activation of weights and the sequence of multiple operations BN, ReLU and convolution. The resulting sub-block is shown in Fig. 4 which we refer to as the *Dense Learning* block.

Our proposed GeneXNet block is finally formed by merging together these two sub-blocks as shown in Fig. 2, which represents a combination of dense learning and residual learning layers. We define several parameters to control the variation of the network design and size across different gene expression data sets. The parameter θ_k controls the number of filters used in the convolution layers. The two parameters θ_D and θ_R define the percentage of dense and residual sub-blocks in the network, where $0 \leq \theta_D \leq 1$ and $0 \leq \theta_R \leq 1$.

The full Gene eXpression Network (GeneXNet) architecture is shown in Fig. 2. It is implemented by feeding the gene expression input volume to multiple layers of GeneXNet blocks each containing a combination of dense and residual learning layers as described above. The network ends with a global average pooling [43] after the last GeneXNet block and a fully connected softmax layer for classification. We experiment with different network sizes having two to four GeneXNet blocks and with different θ_k, θ_D,

TABLE 1. Gene eXpression Network detailed architecture. (implementing a network with 4 blocks, $\theta_k = 32$, $\theta_D = 1$, $\theta_R = 1$).

GeneXNet Block (l)	Output Size	Dense Sub-block		Residual Sub-block	
		Layer operations $\theta_D = 1$	Filters $\theta_k = 32$	Layer operations $\theta_R = 1$	Filters $\theta_k = 32$
Input	(142,142,3)				
Pre-layers	(71,71,64)	$Conv(7x7, 64)$			
GeneXNet Block 1	(36,36,256)	$\begin{bmatrix} Conv(1x1, 128) \\ Conv(3x3, 32) \end{bmatrix} * 6$	$4\theta_k$ θ_k	$\begin{bmatrix} Conv(1x1, 64) \\ Conv(3x3, 64) \\ Conv(1x1, 256) \end{bmatrix} * 2$	$2^l\theta_k$ $2^l\theta_k$ $2^{l+2}\theta_k$
GeneXNet Block 2	(18,18,512)	$\begin{bmatrix} Conv(1x1, 128) \\ Conv(3x3, 32) \end{bmatrix} * 12$	$4\theta_k$ θ_k	$\begin{bmatrix} Conv(1x1, 128) \\ Conv(3x3, 128) \\ Conv(1x1, 512) \end{bmatrix} * 2$	$2^l\theta_k$ $2^l\theta_k$ $2^{l+2}\theta_k$
GeneXNet Block 3	(9,9,1024)	$\begin{bmatrix} Conv(1x1, 128) \\ Conv(3x3, 32) \end{bmatrix} * 24$	$4\theta_k$ θ_k	$\begin{bmatrix} Conv(1x1, 256) \\ Conv(3x3, 256) \\ Conv(1x1, 1024) \end{bmatrix} * 2$	$2^l\theta_k$ $2^l\theta_k$ $2^{l+2}\theta_k$
GeneXNet Block 4	(5,5,2048)	$\begin{bmatrix} Conv(1x1, 128) \\ Conv(3x3, 32) \end{bmatrix} * 16$	$4\theta_k$ θ_k	$\begin{bmatrix} Conv(1x1, 512) \\ Conv(3x3, 512) \\ Conv(1x1, 2048) \end{bmatrix} * 2$	$2^l\theta_k$ $2^l\theta_k$ $2^{l+2}\theta_k$
Classification	(1,1,2048)	*Global Average Pooling*			
	(C-Classes)	*Fully connected (C-Tumor Types) – Softmax*			

θ_R configurations. A detailed architecture is shown in table 1 implementing a network with 4 GeneXNet blocks, $\theta_k = 32$ and both θ_D, θ_R set to 1.

Our results have demonstrated that our proposed network which combines both dense and residual learning layers, has allowed training deeper network architectures with complex data such as gene expressions, despite the large number of genes. The dense layers allow the network to efficiently extract the genetic signatures from multiple tumors and across multiple cancer types. This is achieved by means of re-using the gene expression feature maps learned by different layers, which increases the variation of input signals fed to subsequent layers since it represents the collective knowledge of the network [33]. The residual layers with identity mappings contribute to providing a direct path for information propagation in the forward and backward passes [46] while the connectivity of the dense layers provide each layer with more direct access to the gradients from the loss function and the original input signal [33].

IV. TRANSFER LEARNING USING GENOMIC SIGNATURES OF MULTIPLE CANCER TUMOR TYPES

Our approach for building a comprehensive multi-tissue cancer classifier is by designing the Gene eXpression Network (GeneXNet) with the capability of learning the genomic signatures of whole-transcriptome gene expressions shared across multiple cancer tumor types. By training the model with samples from multiple tissue types collected from multiple sites, the classifier is able to learn and extract complex patterns from the gene expressions that represent genomic and transcriptomic alterations. This allows the classifier to more accurately classify cancer tumors which are resulting from DNA or RNA changes that alter cell behavior across multiple tissues and cause uncontrollable growth and malignancy.

A major advantage is that we can reuse the genomic signatures learned by the trained model to perform very efficient transfer learning to solve one of the biggest challenges in cancer classification which is lack of patient samples. We demonstrate how transfer learning can be used to build and finetune classifiers for other different types of cancer tumors not included in the underlying dataset, which might be lacking sufficient patient samples to be trained independently. By reusing the weights of the pretrained GeneXNet model, we demonstrate how the same network or an extended version of it can be used for feature extraction on a different cancer tumor type. The intuition behind transfer learning comes from recent studies which have performed an integrated multiplatform analysis across multiple cancer types that have revealed similar molecular classification within and across tissues of origin [4], [5]. This means that the discriminative molecular features for one cancer classifier will most likely be relevant for other cancer types. Our pretrained model will have already learned the complex types of genetic alterations and genomic signatures collected from multiple cancer tissue types originating from different organs, and can effectively function as a generic model for cancer classification.

V. VISUALIZING GENOMIC RELATIONSHIPS OF GENE EXPRESSIONS ACROSS MULTIPLE TUMOR TYPES

One of the challenges in using deep learning for disease diagnosis, is that deep networks are conceived as ‘‘black boxes’’ without much interpretation on how these complex models make their decisions [42]. Extensive work has been done to introduce novel visualization techniques for deep networks to help understand and interpret their record breaking performance in computer vision tasks [41], [42], [51]. The output from these techniques can be interpreted by non-experts when studied in conjunction with image or video datasets because they are visually comprehensible. Unfortunately, these methods are not directly applicable to genomic datasets such as gene expressions, since they cannot be visually rendered in a human-friendly form that allows easy interpretations. Our learning system architecture can contribute in solving this problem, since it is designed to address the complex nature of gene expressions.

We introduce two visualization procedures to provide more biological insight on how our proposed deep network is performing cancer classification across multiple tumor types. Our methods are inspired from the work used to visualize intermediate feature activations for CNNs used in image classification [51]. We also build on the methods for Class Activation Maps (CAM) [41], [42] which visualize heatmaps of class activations for deep networks used in image classification and captioning.

A. VISUALIZING CLASS-DISCRIMINATIVE LOCALIZATION MAPS OF GENE EXPRESSIONS

We introduce a visualization method which uses the gradient information flowing into the last convolutional layer of the GeneXNet model to produce gene localization maps highlighting the important regions in the gene expressions which influenced the resulting tumor class prediction. The gene expression data used to train the network is sparse and very high in dimensionality since it represents a snapshot of the whole transcriptome rather than a predetermined subset of genes. By identifying a class-discriminative localization map in the gene expressions, we can identify the subset of genes driving cancer progression and resulted in the model's tumor class prediction. We refer to this localization map as a Gene-Class-Activation-Map (Gene-CAM). For each tumor type, the Gene-CAM is a representation of the discriminative genes used by the network to correctly classify the tumor. The procedure can be summarized as follows:

For a GeneXNet with L blocks, the network will produce a set of intermediate activation feature maps as the output of each block. Let F_l represent the output feature maps of the l^{th} block with dimensions (width: X_l, height: Y_l, depth: D_l). This volume represents the molecular features learned by the network that will be activated when matched with similar patterns in the input gene expressions of a given tumor sample.

Let $f_L^k\,(i,j)$ represent the k^{th} feature map for the last block at special location (i,j). Since the network uses Global Average Pooling (GAP) [43] before the final softmax layer to calculate the spatial average of the feature maps, then the classification score s^c for tumor type c which is used as input to the softmax can be written as:

$$s^c = \sum_k^{D_L} w_k^c \sum_i^{X_L} \sum_j^{Y_L} f_L^k\,(i,j) \tag{7}$$

where c is the tumor class, w_k^c represents the weights for class c with respective to feature map k and D_L is the number of feature maps in the last block before the GAP layer each with width X_L and height Y_L.

We redefine the weights of each feature map with respect to a class as α_k^c by computing the gradient of the score of each class with respect to each feature map as follows:

$$\alpha_k^c = \frac{1}{X_L.Y_L} \sum_i^{X_L} \sum_j^{Y_L} \frac{\partial s^c}{\partial f_{ij}^k} \tag{8}$$

where the new weights α_k^c represent the importance of each feature map for class discrimination. The Gene-Class-Activation-Map (Gene-CAM) is then calculated as:

$$Gene\_CAM^c\,(i,j) = ReLU\left[\sum_k^{D_L} \alpha_k^c \cdot f^k\,(i,j)\right] \tag{9}$$

The resulting map with dimensions (X_L, Y_L) represents a gene localization for the given tumor sample that captures the discriminative regions in the gene expression input matrix which influenced the prediction of the tumor class. The ReLU [38] is applied to obtain only the features that have a positive contribution to the correct class [42].

Finally, to visualize the Gene-CAM we resize it using up-sampling and overlay it against the input gene expression matrix. The resulting heatmap highlights the important regions in the gene expression input matrix which in turn helps identify the subset of genes that are possibly influencing the Cancer tumor and resulted in the model's class prediction.

B. VISUALIZING MOLECULAR CLUSTERS OF INTERMEDIATE FEATURE MAPS

We introduce a visualization procedure for observing the evolution of molecular clusters formed by intermediate gene expression feature maps learned by the network. The genetic signatures learned by the feature maps in the deep layers make the network capable of representing complex genetic alterations shared by tumors across different tissue types. Visualizing the molecular clusters of gene expressions provides more insight on how the network is learning small meaningful relationships between the genes which in turn describe the characteristic influencing the cancer tumor. We demonstrate how this visualization provides the opportunity to study the genomic relationships of gene expressions across multiple tissue types. This is motivated by recent studies which have performed an integrated multi-platform analysis across multiple cancer types that have revealed molecular classification within and across tissues of origin [4], [5].

As in the previous section, for a GeneXNet with L blocks, let F_l represent the output feature maps of the l^{th} block. Let $f_l^k\,(i,j)$ represent the k^{th} feature map for the l^{th} block at special location (i,j). We apply Global Average Pooling (GAP) [43] to each of the intermediate feature maps after each block to convert the volume F_l into a 1-dimensional feature vector F'_l with dimensions $(1, 1, D_l)$ as follows:

$$f_l'^k\,(i,j) = \frac{1}{X_l.Y_l} \sum_i^{X_l} \sum_j^{Y_l} f_l^k\,(i,j) \tag{10}$$

$$F'_l = [f_l'^k\,(i,j)] \quad \forall k \in \{1, .., D_l\} \tag{11}$$

where D_l is the number of feature maps in the l^{th} block each with width X_l and height Y_l. The feature vector F'_l represents the spatial average of the feature maps produced by each filter in the convolutional layer. The intuition behind using GAP is due to its ability to produce a generic localizable deep representation of the features which can be used for class discrimination [42].

We stack together all the feature vectors at the l^{th} block across all N tumor samples to produce what we refer to as a *Gene Feature Map* $(Gene\_Map_l)$ of dimensions (D_l, N):

$$Gene\_Map_l = [F'_l(n)^T] \quad \forall n \in \{1, .., N\} \tag{12}$$

The resulting matrix stores the collective class-discriminative localization maps for the gene expressions at the l^{th} block across all the tumor types. It also represents the collective genetic signatures learned by the feature maps shared by tumors across different organ sites.

Finally, we perform a consensus hierarchical clustering [55] of the gene feature map $Gene\_Map_l$ to generate a $Gene\_Cluster\_Map_l$ which is a molecular clustering that groups each of the tumor types together based on the class discriminative gene localizations extracted from the gene expressions. Consensus clustering is specifically tailored for gene expression data and is based on resampling to reach a consensus across multiple runs of a clustering algorithm and assess the stability of the discovered clusters [55].

By visualizing a heatmap of the resulting clusters, we can observe the evolution of molecular clusters formed by intermediate gene expression feature maps learned by the network. Visualizing the molecular clustering helps in revealing the genomic relationships and high-level structures of gene expressions across multiple cancer tumor types that appeared influential in the cancer tumor progression beyond the standard grouping by anatomical organ site.

The results of applying the visualization procedures to the underlying dataset are described in the experiments.

VI. EXPERIMENTS

A. DATASETS

Our objective was to design a comprehensive multi-tissue cancer classifier capable of detecting complex types of genetic alterations, by learning the genomic signatures of whole-transcriptome wide gene expressions across multiple cancer tissue types. To achieve this objective, the datasets we selected for our experiments included a total of 11,093 human samples for mRNA gene expression quantification, which were collected from 26 different human anatomical organ sites and covering 33 different cancer tumor types. The datasets were obtained from ''The Cancer Genome Atlas'' (TCGA) [30] and generated by means of Total RNA sequencing [4]. Each individual human sample represents the whole transcriptome and includes a total of 60,483 genes annotated against a reference genome. The patients included males and females and the biospecimens were collected from tumor tissue, adjacent normal tissue and normal whole blood samples [4]. Table 2 shows a listing of the 33 cancer tumor types we used in our experiments together with the associated human organ sites and the number of human samples available for each tumor type. One of the biggest challenges in using this dataset is the very small number of human samples in each of the tumor types, compared to the very large number of genes. Most of the tumors only have several hundred samples and some even have less than a hundred samples while we have a total of 60,483 genes for each sample. We represent the gene expression data in a format that makes it suitable as input to our model. We convert each sample into an equivalent 3D volume of genes with dimensions (142, 142, 3). The full dataset for all the 11,093 samples is represented by a 4D input matrix of real numbers with dimensions (11903, 142, 142, 3).

B. CLASSIFICATION EXPERIMENTS

Our experiments demonstrate how the design of our GeneXNet model can be used as a general end-to-end learning system for classification across multiple cancer tissue types without performing the prerequisite process of gene feature selection. We also demonstrate how our model can specifically target the complex nature of whole-transcriptome gene expressions and address the lack of training samples, without suffering from severe overfitting in comparison to using the current state-of-the-art deep CNN models.

We perform several different multi-class and binary classification tasks. For binary classification we predict whether the given sample represents a tumor versus a normal tissue. For multi-class classification we predict for a given sample the type of cancer tumor within each site of origin. The following is an outline of the experiments:

1) We build a *multi-tissue multi-class* classifier by training our model using *ALL* the data which includes 26 organ sites covering 33 tumor types.
2) We build a *multi-tumor* binary classifier for individual organ sites that relatively had the greatest number of samples which included 11 sites as shown in table 3.
3) We repeat the second experiment, but this time we perform *transfer learning* using the weights of the pre-trained model from the first experiment. The objective was to compare the performance between transfer learning using a pre-trained model and full training.
4) We use *transfer learning* to build binary classifiers for organ sites that did not have sufficient data to be trained independently. These included Bile Duct and Esophagus which only had 45 and 147 samples respectively.

C. TRAINING, OPTIMIZATION AND EVALUATION

We use stratified random sampling to divide our datasets into 85% for training/validation and 15% for final testing. We train all models using stratified k-fold cross-validation experimenting with different fold sizes. We use the validation data to optimize the hyperparameters of our models while the test data is strictly used only once as an independent dataset to evaluate the final performance.

Training a deep multi-layer CNN architecture like GeneXNet is a very complex optimization problem as it involves non-convex loss functions [53]. Among the challenges we faced in model optimization is the very high dimensional landscape of the network weight space resulting from training the network with the whole-transcriptome wide gene expressions for every tumor sample. To overcome these problems, we train our model using mini-batch Stochastic Gradient Descent (SDG) with an adaptive learning rate optimization algorithm [48]. We experiment with Adam [39], AdaGrad [40] and RMSprop [48]. We start with a learning rate of $1\mathrm{e}^{-4}$ and divide it by half when the validation loss plateaus for more than 50 epochs. We evaluate the performance of our GeneXNet model with different architectures and sizes by tuning the parameters θ_D, θ_R with values (0, 0.25, 0.5 and 1) and θ_k with values (32, 64). These parameters define the percentage of dense and residual sub-blocks in the network and the number of filters used in the convolution layers.

We evaluate the classification performance of our GeneXNet models using the receiver operating characteristics (ROC) curves [54]. For all experiments, we report the average classification accuracy and ROC area under the curve (AUC) on the Test dataset. The ROC AUC has an advantage of being less sensitive to changes in class distribution as it summarizes the performance over a range of tradeoffs between the true positive and false positive rates [54]. To overcome any potential impact on the classification performance due to class imbalance, we experimented with two different methods for addressing class imbalance. We used Synthetic Minority Over-sampling [59] and Adaptive Synthetic Sampling [60].

We also evaluate the performance of our model in comparison with some of the current state-of-the-art CNN models specifically designed for computer vision tasks. We perform the same multi-class classification task using all the data but replacing our model with the publicly available implementations of ResNet [36], [46], DenseNet [33], NasNet [32] and MobileNet [31], [35].

TABLE 2. Results of multi-tissue classification using 26 organ sites covering 33 tumor types.

Organ Site	Cancer Tumor Type(s)	Total Samples	Accuracy (%)
Adrenal Gland	Adrenocortical carcinoma (ACC), Pheochromocytoma and Paraganglioma (PCPG)	265	100
Bile Duct	Cholangiocarcinoma (CHOL)	45	100
Bladder	Bladder Urothelial Carcinoma (BLCA)	433	98.46
Bone Marrow	Acute Myeloid Leukemia (LAML)	151	91.3
Brain	Glioblastoma multiforme (GBM), Brain Lower Grade Glioma (LGG)	703	100
Breast	Breast invasive carcinoma (BRCA)	1222	99.46
Cervix	Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC)	309	100
Colorectal	Colon adenocarcinoma (COAD), Rectum adenocarcinoma (READ)	698	99.05
Esophagus	Esophageal carcinoma (ESCA)	173	96.15
Eye	Uveal Melanoma (UVM)	80	100
Head and Neck	Head and Neck squamous cell carcinoma (HNSC)	546	100
Kidney	Kidney Chromophobe (KICH), Kidney renal clear cell carcinoma (KIRC), Kidney renal papillary cell carcinoma (KIRP),	1021	99.35
Liver	Liver hepatocellular carcinoma (LIHC)	424	98.44
Lung	Lung adenocarcinoma (LUAD), Lung squamous cell carcinoma (LUSC)	1145	99.42
Lymph Nodes	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma (DLBC)	48	87.5
Ovary	Ovarian serous cystadenocarcinoma (OV)	379	98.25
Pancreas	Pancreatic adenocarcinoma (PAAD)	182	96.43
Pleura	Mesothelioma (MESO)	86	100
Prostate	Prostate adenocarcinoma (PRAD)	551	97.59
Skin	Skin Cutaneous Melanoma (SKCM)	472	98.59
Soft Tissue	Sarcoma (SARC)	265	100
Stomach	Stomach adenocarcinoma (STAD)	407	98.39
Testis	Testicular Germ Cell Tumors (TGCT)	156	100
Thymus	Thymoma (THYM)	121	100
Thyroid	Thyroid carcinoma (THCA)	568	97.67
Uterus	Uterine Corpus Endometrial Carcinoma (UCEC), Uterine Carcinosarcoma (UCS)	643	100
(ALL Sites)	(All Tumors)	11,093	98.93

TABLE 3. Results of multi-tumor binary classification for 11 individual organ sites.

Organ Site	Full Training		Transfer Learning & Finetuning	
	Accuracy (%)	ROC AUC	Accuracy (%)	ROC AUC
Bladder	96.92	1.0	95.38	0.99
Breast	98.37	0.998	98.37	1.0
Colorectal	100	1.0	100	1.0
Head & Neck	98.78	0.985	92.68	1.0
Kidney	100	1.0	100	0.97
Liver	100	1.0	98.44	1.0
Lung	99.42	1.0	99.42	0.94
Prostate	97.59	0.961	97.59	0.94
Stomach	96.77	0.979	96.77	0.88
Thyroid	95.35	0.981	93.02	1.0
Uterus	100	1.0	100	0.89
Bile Duct*	-	-	85.71	0.89
Esophagus*	-	-	92.31	0.99

D. RESULTS

The results of the first experiment which performed *multi-class* classification using *ALL* the data including 26 organ sites covering 33 tumor types are shown in table 2. Our GeneXNet model was able to achieve excellent results with an overall classification accuracy of 98.93% and a ROC AUC of 0.99 on the test dataset. The results show that our model achieved 100% accuracy on 14 different tumor types, even for some tumor types which had very little human samples such as: Bile Duct (CHOL), Eye (UVM) and Pleura (MESO) which only had 45, 80 and 86 samples respectively.

The results of the second experiment which performed *binary* classification for 11 selected individual organ sites are shown in table 3. Our GeneXNet model was able to achieve 100% accuracy for 8 different tumor types and between 95.35% to 99.42% accuracy for the remaining tumors.

The results of the third and fourth experiments which performed *transfer learning* are also shown in table 3. The results show that transfer learning achieved excellent results which are comparable to those achieved using full training. Transfer learning was able to solve the problem for tumor sites such as Bile Duct and Esophagus which did not have sufficient data to be trained independently. By finetuning the pre-trained model, we were able to achieve 92.31% accuracy for Esophagus and 85.71% accuracy for Bile Duct despite that these sites only had 147 and 45 samples respectively.

The results of transfer learning have demonstrated how our pre-trained model was able to effectively function as a generic model for cancer classification. The comprehensive genomic signatures learned by our network allowed performing very efficient transfer learning to solve one of the biggest challenges in cancer classification which is lack of patient samples. We demonstrated how transfer learning can be used to build classifiers for cancer tumors which are lacking sufficient patient samples to be trained independently.

The results for evaluating the performance of our GeneXNet model in comparison with state-of-the-art CNN models is shown in table 4. A comparison between the ROC curves for the different models is shown in Fig. 5. These results demonstrate that our GeneXNet model consistently

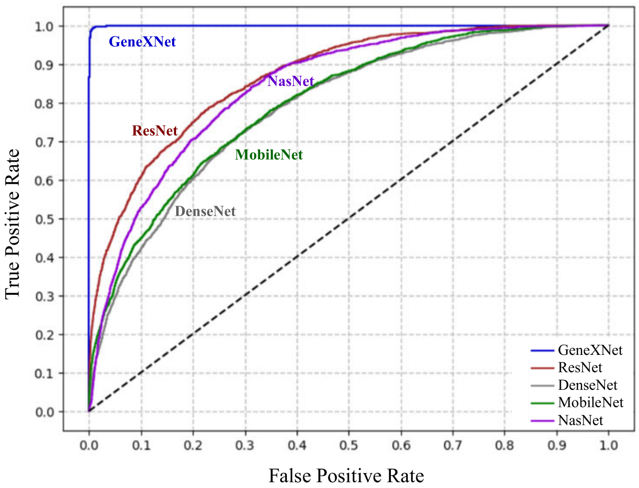

FIGURE 5. Comparison of ROC curves for multi-tissue classification between our GeneXNet model and state-of-the-art CNN models. Our model produced a much higher ROC curve and outperformed other models by a large margin.

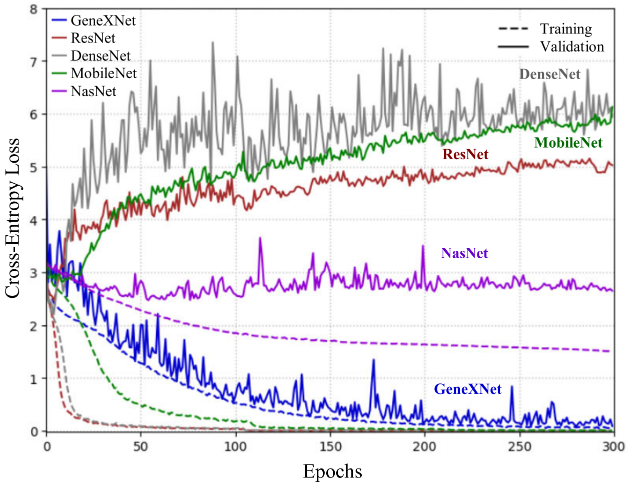

FIGURE 6. Comparison of training and validation cross-entropy Loss for multi-tissue classification, between GeneXNet and other models. Our model achieved minimum loss while other models suffered severe overfitting. Dashed curves are training and solid are validation.

outperformed other CNN models by a large margin. The classification accuracy achieved by our model is 98.93% which is significantly higher than the other models which achieve an accuracy below 37%. Fig. 5 shows that our model produced a much higher ROC curve in comparison to the other models. To provide more insight on this degradation in performance for state-of-the-art models, Fig. 6 shows a comparison between the training and validation curves for each model by plotting the cross-entropy loss across the training epochs. Fig. 6 demonstrates that training these state-of-the-art models which were specifically designed for computer vision tasks, suffered from severe overfitting when presented with the underlying dataset that includes whole transcriptome gene expressions from multiple tumors types.

On the other hand, our GeneXNet model was able to achieve high accuracy in multi-tumor classification while avoiding overfitting. This ability is attributed to the architecture of our model that is designed specifically to target the complex nature of gene expressions and which incorporates both dense and residual learning layers that perform a regularizing effect which allows the network to overcome overfitting.

Our experiments have demonstrated how our model can be used for classification across multiple cancer tissue types without performing the prerequisite gene feature selection. Our model has allowed training deeper network architectures with complex data like whole-transcriptome gene expressions, despite the large number of genes. Our GeneXNet managed to address the lack of training samples, without suffering from severe overfitting in comparison to other CNN models.

The experiments demonstrated that our model design which combines both dense and residual learning layers, helps avoid overfitting and degradation in performance as the network depth increases. The dense layers provide more direct access to the gradients from the loss function and the input, while the residual layers with identity mappings provide a direct path for information propagation in the forward and backward passes.

TABLE 4. Classification performance of GeneXNet in comparison with state-of-the-art CNN models.

Network Model	Accuracy (%)	ROC AUC	Cross Entropy Loss
GeneXNet	**98.93**	**0.99**	**0.06**
ResNet-50 v2 [36]	36.96	0.86	4.9
DenseNet-121 [33]	22.33	0.79	6.09
NasNetMobile [32]	21.61	0.84	2.58
MobileNet v2 [31]	24.96	0.8	5.99

E. RESULTS FOR VISUALIZING CLASS-DISCRIMINATIVE LOCALIZATION MAPS

We apply the visualization procedure to identify a class discriminative Gene-Class-Activation-Map (Gene-CAM) to the underlying dataset to produce a Gene-CAM for each of the 33 individual tumors and then visualize them using heatmaps. Fig. 7 shows the resulting heatmaps of four selected tumor types (Breast, Liver, Stomach and Uterus). By mapping the resulting Gene-CAM to each input sample the network was able to identify a subset of 75 discriminative genes. For visualization, we apply a threshold where each heatmap shows the top 20 genes influencing the underlying tumor across 20 random samples. The rows represent genes, the columns represent samples and the values are the gene expression levels. The gene symbols are displayed on the right of each row with the percentage of samples that have also identified this gene in their Gene-CAM. Each map is a visual representation of the discriminative genes used by the network to correctly classify the tumor.

The strength of our method is that the network automatically identified a small subset of class-discriminative genes out of the total 60,483 genes originally included in each

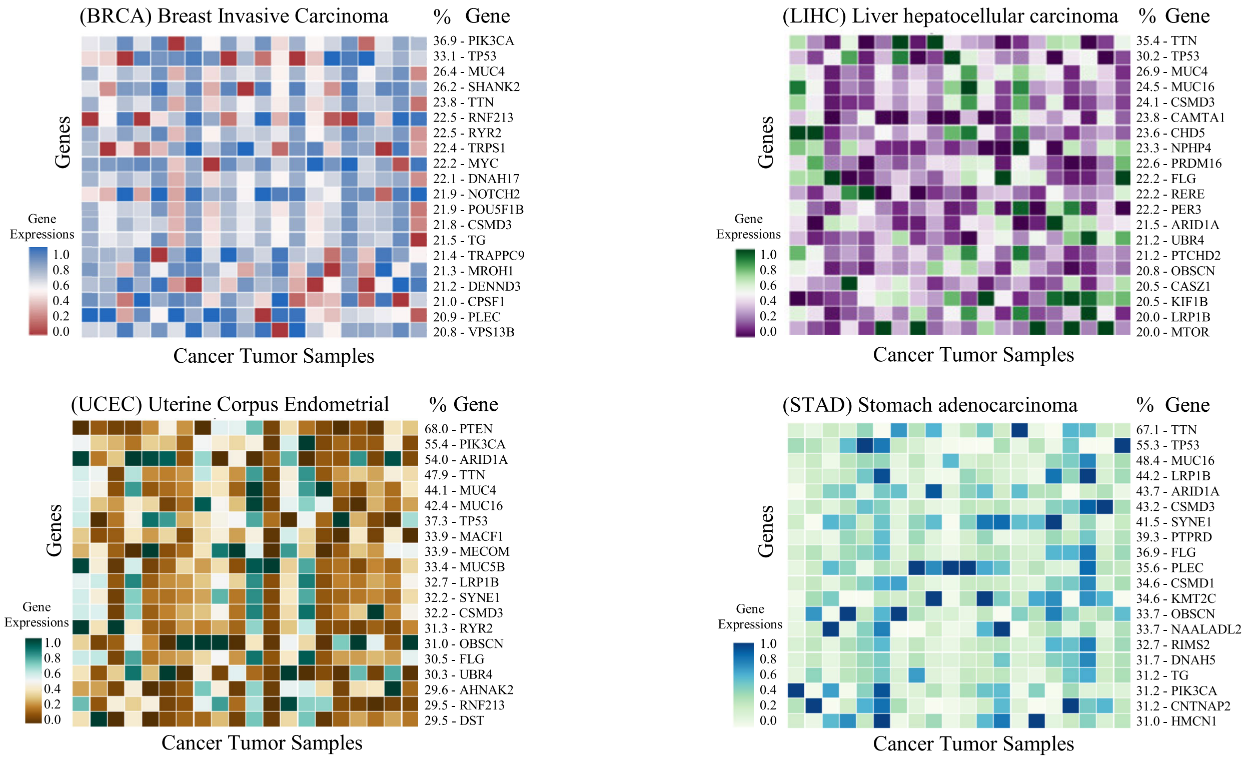

FIGURE 7. Visualizing class-discriminative localization maps highlighting the important regions in the gene expressions which influenced the tumor class prediction. Each map shows the top 20 genes across 20 random samples and is a visual representation of the discriminative genes used by our network to correctly classify the tumor. The rows represent genes, columns represent samples and the values are the gene expression levels.

individual sample. What was also very interesting is that the network automatically identified the TP53 gene as one of the top features common across all tumor types. This result implicitly validates our procedure since TP53 is considered the most commonly mutated gene in all cancers which produces a protein that suppresses the growth of tumors [3].

We also observed from our experiments that some of the identified discriminative genes were also common in at least 30% of samples across different tumor types even though the tissues belonged to different organ sites. This subset includes: TP53, TTN, MUC16, LRP1B, CSMD3, PIK3CA, MUC4, RYR2, USH2A, FLG, PTPRD, CSMD1. These discriminative genes identified by the network have great biological significance for early cancer diagnosis. For example, the mutations of PIK3CA gene are one of the most common in Breast cancer and are reported in over one third of cases [62]. Mutations in TTN gene are associated with one of the most common inherited cardiac disorders known as Hypertrophic Cardiomyopathy (HCM) [61]. MUC16 has a biological role in the progression of Ovarian tumors and there has been substantial work to develop therapeutic approaches to eradicate Ovarian tumors by targeting MUC16 [63]. LRP1B is frequently mutated in Melanoma, Non-small Cell Lung cancer (NSCLC) and other types of tumors. LRP1B is also a potential contributor to the emergence of chemotherapy resistance while treating cancer patients [64]. CSMD3 was identified as the second most frequently mutated gene in Lung cancer after TP53 [3]. MUC4 is a membrane bound mucin gene responsible for progression of several cancers due to its anti-adhesive properties including Bile Duct, Breast, Colon, Esophagus, Ovary, Lung, Prostate, Stomach and Pancreas [61]. Mutations of RYR2 gene are a common cause of abnormal heart failures such as Catecholaminergic Polymorphic Ventricular Tachycardia (CPVT) [61]. PTPRD is frequently mutated in various types of cancer, including Glioblastoma, Melanoma, Breast and Colon [3]. CSMD1 has been found as a tumor suppressor in the development of Breast cancer [61].

F. RESULTS FOR VISUALIZING MOLECULAR CLUSTERS OF INTERMEDIATE FEATURE MAPS

We apply the visualization procedure for observing the molecular clusters formed by intermediate gene expression feature maps to the underlying dataset. We use a GeneXNet with four blocks to produce a molecular clustering of the gene feature maps (*Gene_Cluster_Maps*) after each block. Each *Gene_Cluster_Map* represents a molecular clustering that groups the tumors by organ site based on the class discriminative gene localizations extracted from the gene expressions and learned by the network after each block.

Fig. 8 shows a heatmap of the *Gene_Cluster_Map* for the last block filtered for clusters with at least 200 samples per cluster, which resulted in a total of 17 cluster groups

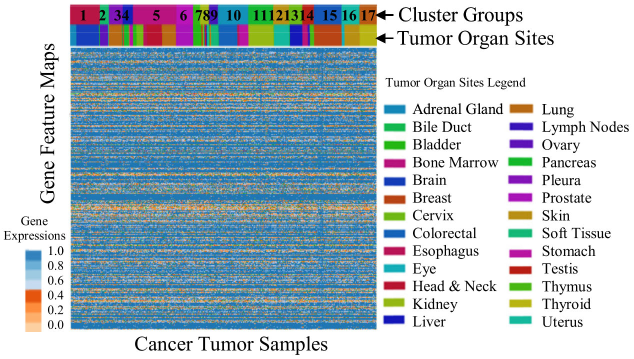

FIGURE 8. Visualizing molecular clusters of intermediate feature maps to reveal genomic relationships across multiple tumors that appeared influential in cancer progression. The heatmap shows a Gene Cluster Map of 17 cluster groups comprising 33 tumors across 26 organ sites.

comprising the 26 organ sites. The rows represent gene localization feature maps, the columns represent samples and the values are the activations of feature maps. The heatmap visually illustrates the genomic relationships and high-level structures of the cancer tumor types across the different organ sites. We observed from our experiments that the number of cluster groups learned by the network in the Gene Cluster Maps decreases as we move towards the deep layers in the network. The feature maps generated after the first block seem to have little in common across the different tumors which is evident by the very large number of resulting cluster groups. As we reach the final network block, we observed that the *Gene_Cluster_Map* has less number of clusters where more clusters have merged together to finally reach only 17 cluster groups. These results have great significance since they demonstrate that as we go deeper in the network, the gene feature maps become more abstract in the sense that they are less representative of the individual tumor samples and more representative of the tumor classes.

We further analyzed the resulting cluster groups in terms of membership of tumor organ sites among the groups as in [4]. We observed that although tissue organ site was mostly a dominant factor for cluster formation, but some clusters also included tumor types across multiple different organs. We also observed that clusters were formed for tumors which appeared to have similar organs or tissue characteristics. For example, Bile Duct and Liver tumors clustered together including CHOL and LIHC. Brain and Nervous system tumors clustered together including LGG and GBM. Kidney and Adrenal Gland tumors formed multiple clusters including KICH, KIRC, KIRP and ACC. Lymph Nodes and Bone Marrow tumors clustered together including DLBC and LAML. Many small overlapping clusters formed together for Stomach, Colorectal, Esophagus and Pancreas tumors including STAD, COAD, READ, ESCA and PAAD. Finally, the remaining clusters were dominated by mostly tumors of a single organ but also included less than 5% of other tumors. These results are very much inline with the molecular characteristics of the underlying dataset reported in [4].

Visualizing the evolution of molecular clusters formed by intermediate gene feature maps, has demonstrated how our proposed GeneXNet is functioning as a comprehensive multi-tumor cancer classifier. The network was capable of learning the complex molecular signatures and genetic alterations shared by tumors across different tissue types and organ sites. This also demonstrates how the network was able to perform efficient transfer learning by using the pre-trained models as a generic multi-tumor feature extractor to build additional classifiers for any individual tumor types.

VII. CONCLUSION

We proposed a deep learning framework for cancer diagnosis by developing a multi-tissue cancer classifier based on whole-transcriptome gene expressions. We introduced a new CNN architecture specifically designed to address the complex nature of whole-transcriptome gene expressions and demonstrated how it can be used as a general end-to-end learning system for classification across multiple cancer tissue types without performing the prerequisite process of gene feature selection. We demonstrated how the genetic signatures learned by our model can be used for transfer learning to build classifiers for other types of cancer tumors which are lacking sufficient patient samples to be trained independently. We contributed in providing more confidence in using deep learning for medical diagnosis by introducing visualization procedures to provide biological insight on how our model is performing classification across multiple tumors.

We believe there is great potential for further research to expand on our work for cancer diagnosis. Our work focused on designing a multi-tissue cancer classifier based on Total RNA Sequencing using gene expressions from coding mRNA. Future work can explore learning more complex genomic signatures by including Omics data using other multiple forms of NGS platforms and experimental strategies such as DNA hypermethylation, aneuploidy, non-coding microRNA, DNA Copy Number Variants (CNV) and Reverse Phase Protein Arrays (RPPA). This will provide the opportunity to create a more comprehensive repository of pretrained models readily available for cancer classification using transfer learning. Future work can also target cancer diagnosis and improving classifier performance by designing *Ensemble Models* which could integrate multiple genome-wide platforms by learning molecular signatures across multiple forms of Omics data.

REFERENCES

[1] *World Cancer Report 2019*, Int. Agency Res. Cancer, World Health Org., Geneva, Switzerland, 2019.

[2] World Health Organization. *Cancer Prevention*. [Online]. Available: https://www.who.int/health-topics/cancer

[3] US National Cancer Institute (NCI). *Cancer Research*. [Online]. Available: https://www.cancer.gov/research.

[4] K. A. Hoadley *et al.* "Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer," *Cell*, vol. 173, no. 2, p. 291-304.e6, Apr. 2018.

[5] K. A. Hoadley *et al.*, "Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin," *Cell*, vol. 158, no. 4, pp. 929–944, Aug. 2014.

[6] P. Wu and D. Wang, "Classification of a DNA microarray for diagnosing cancer using a complex network based method," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 3, pp. 801–808, May 2019.

[7] C. Peng, X. Wu, W. Yuan, X. Zhang, and Y. Li, "MGRFE: Multilayer recursive feature elimination based on an embedded genetic algorithm for cancer classification," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Jun. 10, 2019, doi: 10.1109/TCBB.2019.2921961.

[8] F. Hu, Y. Zhou, Q. Wang, Z. Yang, Y. Shi, and Q. Chi, "Gene expression classification of lung adenocarcinoma into molecular subtypes," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Mar. 18, 2019, doi: 10.1109/TCBB.2019.2905553.

[9] H. Lu, H. Gao, M. Ye, and X. Wang, "A hybrid ensemble algorithm combining AdaBoost and genetic algorithm for cancer classification with gene expression data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Nov. 8, 2019, doi: 10.1109/TCBB.2019.2952102.

[10] J. Xu, P. Wu, Y. Chen, Q. Meng, H. Dawood, and M. M. Khan, "A novel deep flexible neural forest model for classification of cancer subtypes based on gene expression data," *IEEE Access*, vol. 7, pp. 22086–22095, 2019.

[11] C.-Q. Xia, K. Han, Y. Qi, Y. Zhang, and D.-J. Yu, "A self-training subspace clustering algorithm under low-rank representation for cancer classification on gene expression data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 4, pp. 1315–1324, Jul. 2018.

[12] E. Razak, F. Yusof, and R. A. Raus, "Classification of miRNA expression data using random forests for cancer diagnosis," in *Proc. Int. Conf. Comput. Commun. Eng. (ICCCE)*, Jul. 2016, pp. 187–190.

[13] S. H. Bouazza, N. Hamdi, A. Zeroual, and K. Auhmani, "Gene-expression-based cancer classification through feature selection with KNN and SVM classifiers," in *Proc. Intell. Syst. Comput. Vis. (ISCV)*, Mar. 2015, pp. 1–6.

[14] S. A. Ludwig, D. Jakobovic, and S. Picek, "Analyzing gene expression data: Fuzzy decision tree algorithm applied to the classification of cancer data," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Aug. 2015, pp. 1–8.

[15] J.-X. Liu, Y. Xu, C.-H. Zheng, H. Kong, and Z.-H. Lai, "RPCA-based tumor classification using gene expression data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 4, pp. 964–970, Jul. 2015.

[16] K. Liu, J. Ye, Y. Yang, L. Shen, and H. Jiang, "A unified model for joint normalization and differential gene expression detection in RNA-seq data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 2, pp. 442–454, Mar. 2019.

[17] J. M. Knight, I. Ivanov, K. Triff, R. S. Chapkin, and E. R. Dougherty, "Detecting multivariate gene interactions in RNA-seq data using optimal Bayesian classification," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 2, pp. 484–493, Mar. 2018.

[18] K. R. Kukurba and S. B. Montgomery, "RNA sequencing and analysis," *Cold Spring Harb Protocols*, vol. 2015, no. 11, Nov. 2015, Art. no. pdb.top084970.

[19] J. E. Dancey, P. L. Bedard, N. Onetto, and T. J. Hudson, "The genetic basis for cancer treatment decisions," *Cell*, vol. 148, no. 3, pp. 409–420, Feb. 2012.

[20] S. Sleijfer, J. Bogaerts, and L. L. Siu, "Designing transformative clinical trials in the cancer genome era," *J. Clin. Oncol.*, vol. 31, no. 15, pp. 1834–1841, May 2013.

[21] L. E. MacConaill, "Existing and emerging technologies for tumor genomic profiling," *J. Clin. Oncol.*, vol. 31, no. 15, pp. 1815–1824, May 2013.

[22] J. M. Rizzo and M. J. Buck, "Key principles and clinical applications of 'next-generation' DNA sequencing," *Cancer Prevention Res.*, vol. 5, no. 7, pp. 887–900, Jul. 2012.

[23] Z. Wang, M. Gerstein, and M. Snyder, "RNA-seq: A revolutionary tool for transcriptomics," *Nature Rev. Genet.*, vol. 10, no. 1, pp. 57–63, Jan. 2009.

[24] C. Liu and H. S. Wong, "Structured penalized logistic regression for gene selection in gene expression data analysis," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 1, pp. 312–321, Jan. 2019.

[25] K. R. Kavitha, A. V. Ram, S. Anandu, S. Karthik, S. Kailas, and N. M. Arjun, "PCA-based gene selection for cancer classification," in *Proc. IEEE Int. Conf. Comput. Intell. Comput. Res. (ICCIC)*, Dec. 2018, pp. 1–4.

[26] S. An, J. Wang, and J. Wei, "Local-nearest-neighbors-based feature weighting for gene selection," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 15, no. 5, pp. 1538–1548, Sep. 2018.

[27] J. C. Ang, A. Mirzal, H. Haron, and H. N. A. Hamed, "Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 5, pp. 971–989, Sep. 2016.

[28] J. Tang and S. Zhou, "A new approach for feature selection from microarray data based on mutual information," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 6, pp. 1004–1015, Nov. 2016.

[29] J.-X. Liu, Y. Xu, Y.-L. Gao, C.-H. Zheng, D. Wang, and Q. Zhu, "A class-information-based sparse component analysis method to identify differentially expressed genes on RNA-seq data," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 13, no. 2, pp. 392–398, Mar. 2016.

[30] *The Cancer Genome Atlas (TCGA) Research Network*. [Online]. Available: https://www.cancer.gov/tcga

[31] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[32] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8697–8710.

[33] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.

[34] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 29, 2019, doi: 10.1109/TPAMI.2019.2913372.

[35] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*. [Online]. Available: https://arxiv.org/abs/1704.04861

[36] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Comput. Vis. ECCV*, Cham, Switzerland, 2016, pp. 630–645.

[37] G. Huang, Y. Sun, Z. Liu, D. Sedra, and K. Q. Weinberger, "Deep Networks with Stochastic Depth," in *Proc. Comput. Vis. ECCV*, Cham, Switzerland, 2016, pp. 646–661.

[38] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Int. Conf. Mach. Learn.*, New York, NY, USA, 2010, pp. 807–814.

[39] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, San Diego, CA, USA, May 2015.

[40] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *J. Mach. Learn. Res.*, vol. 12, pp. 2121–2159, Feb. 2011.

[41] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2921–2929.

[42] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, Feb. 2020.

[43] M. Lin, Q. Chen, and S. Yan, "Network in network," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, Banff, AB, Canada, Apr. 2014.

[44] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, Lille, France, vol. 37, Jul. 2015, pp. 448–456.

[45] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, San Francisco, CA, USA, 2017, pp. 4278–4284.

[46] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[47] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.

[48] S. Ruder, "An overview of gradient descent optimization algorithms," 2016, *arXiv:1609.04747*. [Online]. Available: https://arxiv.org/abs/1609.04747

[49] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.

[50] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556

[51] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Comput. Vis. ECCV*, Cham, Switzerland, 2014, pp. 818–833.

[52] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[53] Y. LeCun, "Learning invariant feature hierarchies," in *Proc. 12th Int. Conf. Comput. Vis.*, Berlin, Germany, 2012, pp. 496–505.

[54] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, Jun. 2006.

[55] M. D. Wilkerson and D. N. Hayes, "ConsensusClusterPlus: A class discovery tool with confidence assessments and item tracking," *Bioinformatics*, vol. 26, no. 12, pp. 1572–1573, Jun. 2010.

[56] L. J. V. Veer *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[57] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, "Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, Oct. 1999.

[58] Y.-H. Zhang, T. Zeng, X. Pan, W. Guo, Z. Gan, Y. Zhang, T. Huang, and Y.-D. Cai, "Screening dys-methylation genes and rules for cancer diagnosis by using the pan-cancer study," *IEEE Access*, vol. 8, pp. 489–501, 2020.

[59] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[60] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IEEE World Congr. Comput. Intell.)*, Jun. 2008, pp. 1322–1328.

[61] US National Library of Medicine, Bethesda (MD). (2004). *Gene: National Center for Biotechnology Information*. [Online]. Available: https://www.ncbi.nlm.nih.gov/gene/

[62] D. Zardavas, W. A. Phillips, and S. Loi, "PIK3CA mutations in breast cancer: Reconciling findings from preclinical and clinical data," *Breast Cancer Res.*, vol. 16, no. 1, p. 201, Feb. 2014.

[63] M. Felder, A. Kapur, J. Gonzalez-Bosquet, S. Horibata, J. Heintz, R. Albrecht, L. Fass, J. Kaur, K. Hu, H. Shojaei, R. J. Whelan, and M. S. Patankar, "MUC16 (CA125): Tumor biomarker to cancer therapy, a work in progress," *Mol. Cancer*, vol. 13, no. 1, p. 129, 2014.

[64] P. A. Cowin *et al.*, "LRP1B deletion in high-grade serous ovarian cancers is associated with acquired chemotherapy resistance to liposomal doxorubicin," *Cancer Res.*, vol. 72, no. 16, pp. 4060–4073, Aug. 2012.

TAREK KHORSHED (Member, IEEE) received the B.S. degree in computer science engineering from Alexandria University, Egypt, in 1996, and the M.S. degree from Middlesex University, London, U.K., in 2011. He is currently pursuing the Ph.D. degree in computer science engineering with The American University in Cairo, Egypt.

From 2003 to 2015, he worked as an Information Technology Officer with the World Health Organization Regional Office, Egypt. Since 2016, he has been working as a Lead Technology Architect with the World Health Organization Headquarters, Geneva, Switzerland. His research interests include artificial intelligence, machine learning, computer vision, and bioinformatics.

Mr. Khorshed was a recipient of the Ph.D. Fellowship in computer science engineering and was awarded a certificate of academic honor for Outstanding Academic Achievement from The American University in Cairo, Egypt. He is a member of the IEEE Computational Intelligence Society.

MOHAMED N. MOUSTAFA (Member, IEEE) received the Ph.D. degree in electrical engineering from The City University of New York, USA, in 2001.

From 1998 to 2017, he was the Principal Research Scientist with IDEMIA, USA, Germany, and France, where he conducted research in machine intelligence in face and iris recognition. He has been an Associate Professor with the Computer Science and Engineering Department, The American University in Cairo, since 2011, where he is currently the Director of the Machine Intelligence Laboratory. He is also the Principal Research Scientist with VKANSEE USA. He holds four issued U.S. patents and coauthored more than 50 research articles published in international journals and conferences in the field of biometrics, computer vision, and machine learning.

Dr. Moustafa is a member of the IEEE Computational Intelligence Society and the IEEE Technical Committee on Pattern Analysis and Machine Intelligence.

AHMED RAFEA served as the Chair of the Computer Science Department and the Vice Dean with the Faculty of Computers and Information, Cairo University. He also served as a Visiting Professor with San Diego State University and National University, USA. He was the Principal Investigator of several projects for developing Intelligent Systems and Machine Translation in collaboration with European and American Universities. He is currently a Computer Science Professor and the Ex-Chair of the Computer Science and Engineering Department, The American University in Cairo. He has led many projects aiming at using Artificial Intelligence and Expert Systems Technologies for the development of the Agriculture sector in Egypt. He has authored over 200 scientific articles in International and National Journals, Conference Proceedings, and Book chapters. His research interests are data, text and web mining, natural language processing and machine translation, knowledge engineering, and knowledge-based system development.

Dr. Rafea was a member with the Center of Excellence on Data Mining and Computer Modeling, sponsored by the Ministry of Communication and Information Technology.