Staged Inference Using Conditional Deep Learning for Energy Efficient Real-Time Smart Diagnosis

Maryam Parsa, Priyadarshini Panda, Shreyas Sen, and Kaushik Roy*

Abstract— **Recent progress in biosensor technology and wearable devices has created a formidable opportunity for remote healthcare monitoring systems as well as real-time diagnosis and disease prevention. The use of data mining techniques is indispensable for analysis of the large pool of data generated by the wearable devices. Deep learning is among the promising methods for analyzing such data for healthcare applications and disease diagnosis. However, the conventional deep neural networks are computationally intensive and it is impractical to use them in real-time diagnosis with low-powered on-body devices. We propose Staged Inference using Conditional Deep Learning (SICDL), as an energy efficient approach for creating healthcare monitoring systems. For smart diagnostics, we observe that all diagnoses are not equally challenging. The proposed approach thus decomposes the diagnoses into preliminary analysis (such as healthy vs unhealthy) and detailed analysis (such as identifying the specific type of cardio disease). The preliminary diagnosis is conducted real-time with a low complexity neural network realized on the resource-constrained on-body device. The detailed diagnosis requires a larger network that is implemented remotely in cloud and is conditionally activated only for detailed diagnosis (unhealthy individuals). We evaluated the proposed approach using available physiological sensor data from Physionet databases, and achieved 38% energy reduction in comparison to the conventional deep learning approach.**

I. INTRODUCTION

With the increase in the world's aging population and the corresponding increase in aggregated costs of hospitalization and patient care, there is an urgent need for providing new methods of health monitoring and diagnosis at affordable cost. The National Health Expenditures Accounts (NHEA), reported by Centers for Medicare and Medicaid Services, show that United States healthcare outlay increased by 5.8% in the year 2015 and reached $3.2 trillion which is $9990 per person. This spending accounts for 17.8% of the United States Gross Domestic Product (GDP) in 2015 [1].

Advancement in biosensor technology, and in turn, wearable devices has dramatically influenced the manner in which people can manage and monitor their health. Users will have a much clearer picture of their health as more physiological data is collected. This not only improves the quality of disease diagnosis, but also increases the chance of disease prevention since detailed health information is known ahead of time. Real-time healthcare monitoring by communicating with wearable devices will play a vital role in early diagnosis and disease prevention.

*All authors are with the School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA.
{mparsa, pandap, shreyas, kaushik}@purdue.edu

To implement real-time diagnosis on any low-power on-body device such as cell-phones or smart watches, energy efficient designs are indispensable. Any development in remote healthcare diagnosis using wearable devices is inseparable from data analysis techniques that are necessary for handling and analyzing the amount of data created from the on-body sensing devices. Data mining, and machine learning techniques, specifically Deep Learning Networks (DLNs), are the major approaches used for healthcare diagnosis based on physiological data [2] - [8]. In fact, deep learning algorithms have already found widespread use in medical imaging applications such as tumor detection from MRI data [9]. Moreover, recent studies show that DLNs predict with a higher accuracy than radiologists [9]. This remarkable performance of DLNs is mainly due to the fact that they extract characteristic features from a given data during the learning process and make predictions based on these important features. While DLNs open up several new possibilities in the field of medical application, the high computational cost incurred with such networks prohibits their usage on low-powered wearable devices for real-time diagnosis.

Interestingly, we make a keen observation that for many healthcare diagnostic applications, all diagnoses are not equally challenging. For instance, identifying whether a person is healthy/sick is much easier than determining the specific cause for the sickness. In fact, the complexity of DLNs (that translates to energy consumption) is directly proportional to the difficulty of the diagnoses. Therefore, on-body devices that can easily realize low-complexity DLNs can conduct simple diagnoses in real-time with low energy consumption. For more detailed diagnosis, a larger DLN can be used to provide more accurate diagnosis.

Based on the above insight, we propose a staged approach for conducting real-time healthcare diagnosis on low-powered devices using deep learning networks. Essentially, we conduct the diagnosis in two stages: on-body and in-cloud. While simple anomaly detection (healthy or sick) is performed through a simple network implemented on an on-body device, the more detailed diagnosis (with specific details of causes underlying the ailment) will be done remotely in the complex DLN in the cloud. In addition to the staged approach, we further utilize a technique called Conditional Deep Learning (CDL) proposed in [10] that allows us to distribute the layers of a DLN between a low-powered device and remote servers in cloud without any loss in prediction accuracy. Typically, in most healthcare diagnostics, the entire DLN (implemented in-cloud) is used for any diagnosis, irrespective of its difficulty. We believe that our proposed staged approach could potentially revolutionize energy-

978-1-5090-2809-2/17/$31.00 ©2017 IEEE

efficient real-time diagnosis, paving the way for wearable on-body devices.

II. STAGED INFERENCE FOR REAL TIME SMART DIAGNOSIS

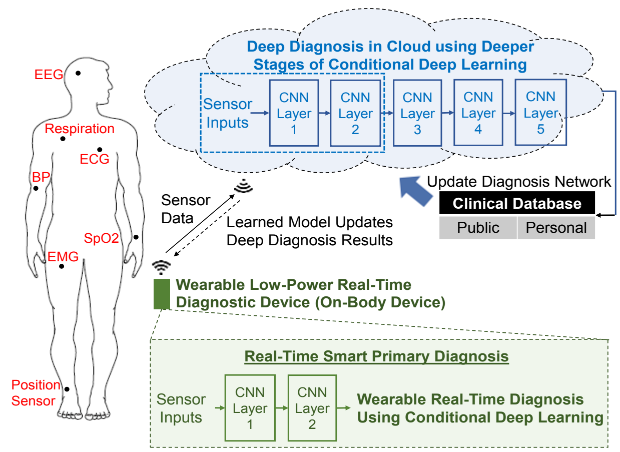

Figure 1. Overview of Real-Time Smart Health Diagnosis System

Fig. 1 shows the conceptual view of our proposed methodology. The DLN (shown with 5 layers) is trained on the available general healthcare data, is available in the cloud, for detailed diagnosis. The on-body diagnostic device on a low-powered platform, processes the sensor inputs directly received from the body to perform preliminary analysis. The on-body diagnostic device is basically the first stage wherein the initial few already-trained layers (Layer 1 and Layer 2) of the DLN are deployed.

The convolutional layers of a DLN generally learn a hierarchy of features that transition from generic (like edges for image applications) to specific (whole object parts) as we go deeper into the network. It is intuitive that the generic features learnt, and hence, the initial convolutional layers can be used to conduct a simplified analysis with high accuracy on an easy classification problem. On the other hand, the deeper layers or specific features need to be used for more difficult analysis to make accurate classifications.

We utilize the usefulness of the convolution layer features to distribute the layers of a network between the on-body device and remote servers in cloud as shown in Fig. 1. The on-body device performs simple diagnosis (such as healthy or sick) utilizing the initial layers of the DLN while specific diagnosis (say, the cause of the sickness) is conducted using the latter layers. Furthermore, we incrementally retrain the DLN in cloud with the personal data (assuming the personal data available from sensor is annotated and labelled at regular intervals) to transform the initial general DLN into a more personalized diagnosis system. It should be noted that retraining is done with the personal as well as the already available general healthcare data to ensure that the DLN does not yield errors due to overfitting. The new weights corresponding to the initial layers (C1, C2) from the retrained network are uploaded onto the device. This process of retraining at regular intervals with personal information makes the diagnostic device more personalized over time.

Although retraining increases the overall runtime complexity of the training process, retraining is still incremental i.e. it is carried out for very few iterations. Typically, the training process in DLNs takes several tens to hundreds of iterations and therefore the increase in run-time complexity due to retraining is small. Also, in typical use cases, the training process is performed once or very infrequently. On the other hand, the testing or evaluation phase, in which the actual diagnostic is performed using the DLN extends for much longer periods of time. The diagnostic device, in reality, needs to process and evaluate the sensor data continuously. Since our proposed approach with distributed layers and staged inference can yield significant energy benefits in the more critical evaluation phase, a small increase in the cost of training is a favorable trade-off.

III. CONDITIONAL DEEP LEARNING METHODOLOGY FOR STAGED INFERENCE

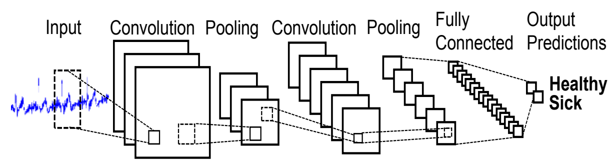

Figure 2. An example of Convolutional Deep Learning Network

Before we discuss the details of our methodology, we briefly review the convolutional deep learning networks (CNNs). Fig. 2 shows a basic structure of a DLN with two convolutional layers followed by pooling layers. Convolutional layers are responsible for convolving a set of weight kernels with a portion of the previous layer to form an array of output maps. The features corresponding to the inputs are learnt by repeating the kernels across the entire input space. Pooling layers are the non-linear down-sampling units that lower the dimesionality of the output maps as we go deeper into the network. Convolutional hierachical organization in deep learning networks enables features to be learnt in a generic-to-specific manner. We exploit this to develop an architecture in which easy classification can be done early without activating the latter layers of the DLN network.

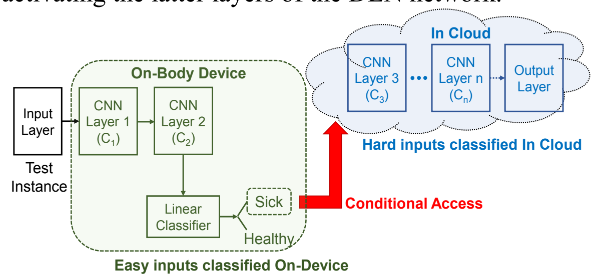

Figure 3. Staged Inference using Conditional Deep Learning (SICDL). Conditional Deep Learning Network with Distributed Layers between On-Body Device and Cloud for Staged Inference

Fig. 3 illustrates our methodology of staged inference using CDL (SICDL) for real-time smart diagnosis in case of a single on-body device. We adopt a similar strategy as proposed in [10] to implement the staged approach. As mentioned in section II, we take a DLN already trained in cloud and then distribute its layers between the on-body device and remote servers in cloud to conduct diagnosis on real time sensor data. To perform a simple classification (healthy vs sick) with the generic features learnt at the initial layers (C1, C2 shown in Fig. 3), we append a linear classifier onto the on-body device. For a given input instance, the output of the C2 layer is fed to the linear classifier that performs the preliminary diagnosis. Depending upon the output of the linear classifier implemented on the on-body device, the latter layers of the DLN from the cloud are remotely accessed. The linear classifier in addition the class label (i.e. healthy or sick) also generates a confidence value associated with each label. This confidence is utilized to decide when the remote layers in the cloud need to be conditionally accessed to perform more specific diagnoses:

- If the linear classifier produces sufficient confidence for the "sick" label, the patient is deemed to be unhealthy and is notified to the remote servers to obtain the exact cause and other details related to the specific ailment.
- If the linear classifier produces sufficient confidence for both healthy and sick class labels, then the input sensor data is deemed to be difficult to classify with the current on-body network and is passed to the cloud for analysis.

It is evident that the cloud or deeper layers are only accessed when the patient is unhealthy. While the illustration in Fig. 3 shows the implementation of our methodology for a single on-body device and cloud, the proposed staged approach with CDL can be extended to several low-powered devices.

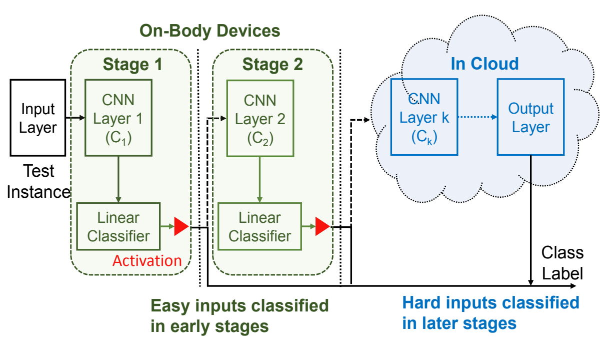

Figure 4. SICDL for More than One On-body Device

Fig. 4 illustrates the generic methodology that extends the staged approach across many on-body devices. The basic idea is to build a sequence of network stages wherein the layers of a network are distributed across several low-powered devices based on the limitations on memory and computational resources posed by the device. Here, each stage denotes that the particular convolutional layer with a linear classifier is implemented for an on-body device. Again, based on the confidence of the output at each linear classifier, the next stage (or cloud) is activated. As an example, stage 1 can do a simple diagnosis of healthy versus sick. If the predicted output from stage 1 is sick, stage 2 is activated. Stage 2 then performs a slightly difficult diagnosis, for instance, identifying whether the sickness, say, is heart related or simply caused by exhaustion. In case the prediction is heart related (say), the cloud is accessed to perform a detailed diagnosis. It is worth mentioning that the diagnosis in case the person is "healthy" is terminated at the first stage itself without activating the latter layers of the DLN. Similarly, for exhaustion (say), the diagnosis is terminated at stage 2. Hence, it is evident that our proposed methodology conditionally activates the layers of the network (or the corresponding on-body devices) based on the difficulty of the diagnosis which leads to a computationally efficient design with reduced energy consumption while yielding competitive accuracy. Furthermore, the cloud is only accessed when a person is sick (as in Fig. 3) or for more detailed disease related diagnosis (as in Fig. 4). In general cases, the fraction of times a person is sick is usually less frequent. Most of the time (excluding all special cases) when a person is healthy, only the smaller network on the on-body device is activated resulting in significant energy-savings compared to the traditional approach where the entire DLN is used for all diagnoses.

IV. Experimental Methodology

As a proof of concept for real-time energy-efficient smart diagnosis (note physiological data from wearable devices were not available), the proposed SICDL approach was examined on a set of available physiological data from Physionet databases [11]. For our experiments, ten different databases were considered to capture physiological records for healthy and sick people with various cardiac and non-cardiac diseases. The ten selected databases from Physionet [11], their descriptions, and the number of recordings from each of them for different categories are summarized in Table 1.

The size of the data that was used in this experiment is not sufficient for any type of deep learning analysis. Therefore, the baseline and SICDL staged approaches were performed on simple feed-forward neural networks to show a proof of concept and energy efficiency of the approach. In the future, the data from the wearable devices will also be added to the available physiological sensory data in literature and the full training step will be performed in a staged inference using conditional deep learning platform.

As shown in Fig. 5, the baseline network (without any stages) is capable of classifying the input instances to five categories with accuracy of 92.8%. However, the staged approach is able to categorize the input instances based on their priority with higher accuracies of 95.5%, 98.8%, and 100% for different stages of healthy/sick, cardiac/non-cardiac and three different types of cardiac diseases categories.

In this experiment, it is assumed that energy efficiency of each approach, is quantified as the average number of Operations Per Input (OPS). Fig. 6a shows energy efficiency comparison of the baseline and the proposed methodology. The healthy people are categorized in the first stage. Consequently, the energy consumption for identifying healthy people is reduced by 75%. The non-cardiac diseases are distinguished in the second stage which reduced the energy usage by half. Different types of cardiac diseases, which are classified at the final stage of the staged approach, each reduces the energy consumption by 25%. The total energy consumption of the proposed staged approach is calculated by summing the energy consumption of each of the three simple networks in different stages. This energy is 38% less than the energy consumption of one complex baseline network. In terms of accuracy, as shown in Fig. 6b, the performance of the staged approach is better than the baseline. This is due to the fact that simple networks with fewer neurons and synapses can be trained easily, and therefore achieve a better least square error as compared to the baseline network.

TABLE I. SUMMARY OF THE SELECTED DATABASES FROM PHYSIONET [11]

No	**Name**	**Abbr.**	**Description**	**Category Name in Experiment (Number of Recordings)**
1	Arterial Fibrillation	AF	Patients with Atrial Fibrillation	Cardiac-AF (30)
2	Cyclic Alternating Pattern	CAP	Sleep disorders	Healthy (14)/ Non-Cardiac (9)
3	Physionet/Computing in Cardiology Challenge	CH15	ICU bedside monitor with life-threatening arrhythmia	Cardiac-CH15 (632)
4	Cerebral Haemodynamic Autoregulatory Information System	CHARIS	Traumatic Brain Injury	Non-Cardiac (13)
5	Stress Recognition in Automobile Drivers	DRIVE	Driving on a predefined route with ECG, EMG, GSR, and respiration recordings	Healthy (16)
6	Sleep-EDF	EDF	Difficulty falling asleep	Healthy (39)/ Non-Cardiac (22)
7	The Fantasia	FANTASIA	ECG and respiration recordings of young and old healthy people	Healthy (34)
8	Physiologic Response to Changes in Posture	PRCP	Database with ECG and ABP signals	Healthy (10)
9	Physikalisch-Technische Bundesanstalt	PTB	Collection of 15-lead ECGs	Healthy (50)/ Cardiac-PTB (212)
10	University College Dublin Sleep Apnea	UCD	Sleep under Temazepam	Non-Cardiac (24)
Number of Total Recordings = 1105 (Total Healthy: 163, Total Sick: 942) (Cardiac: 874 [Cardiac-CH15: 632, Cardiac-PTB: 212, Cardiac-AF: 30], Non-cardiac: 68)				

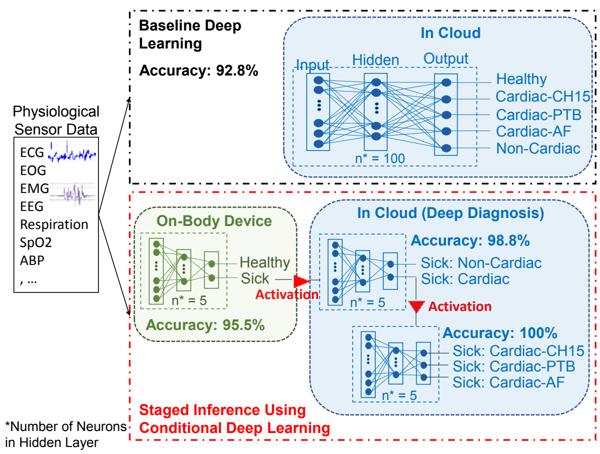

Figure 5. Comparison between the baseline and the SICDL approaches, inputs, outputs, network structures, and accuracies

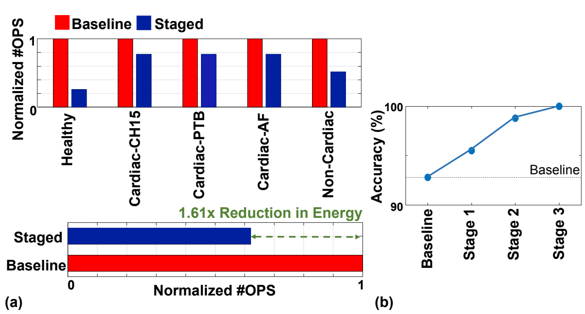

Figure 6. (a) Normalized number of Operations Per Input (OPS) for SICDL with respect to the baseline for different categories. (b) Accuracy comparison between the baseline and different stages of SICDL approach

V. SUMMARY AND CONCLUSION

We developed a technique that enables a real-time smart health diagnosis with an energy efficient staged inferencing using conditional deep learning. This was carried out by identifying the difficulty of the input instances and separating the easy ones from more challenging by adding linear classifiers and activation modules at the output of each convolutional layer. The first stage of the classification corresponds to simpler more energy efficient networks that are performed on wearable low-power on-body devices. In the following stages, the more complicated input instances are diagnosed in the cloud. The proposed staged/conditional approach, in comparison to the baseline approach (without stage), can lead to significant energy reduction while maintaining high accuracy.

VI. ACKNOWLEDGEMENT

The research was funded in part by National Science Foundation, Intel/Semiconductor Research Corporation Education Alliance (SCRCEA) and by Vannevar Bush Fellows program.

REFERENCES

[1] "National Health Expenditures 2015 Highlights. Centers for Medicare & Medicaid Services"

[2] F. H, et al, "Robust medical ad hoc sensor networks (MASN) with wavelet-based ECG data mining," *Ad Hoc Netw.*, vol. 6, no. 7, 986–1012, 2008.

[3] K. H., et al, "Low-energy Formulations of Support Vector Machine Kernel Functions for Biomedical Sensor Applications," *J. Signal Process. Syst.*, vol. 69, no. 3, pp. 339–349, Dec. 2012.

[4] D. G, et al., "Automated diagnosis of Coronary Artery Disease affected patients using LDA, PCA, ICA and Discrete Wavelet Transform," *Knowl.-Based Syst.*, vol. 37, pp. 274–282, Jan. 2013.

[5] T. H. N. V., et al, "Online discovery of Heart Rate Variability patterns in mobile healthcare services," *J. Syst. Softw.*, vol. 83, (10), 1930–1940, 2010.

[6] W. K., et al, "Sleep and Wake Classification With ECG and Respiratory Effort Signals," *IEEE Trans. Biomed. Circuits Syst.*, vol. 3, no. 2, pp. 71–78, Apr. 2009.

[7] V. P., et al, "Decision Trees: An Overview and Their Use in Medicine," *J. Med. Syst.*, vol. 26, no. 5, pp. 445–463, Oct. 2002.

[8] J.-Y. Yeh, et al, "Using data mining techniques to predict hospitalization of hemodialysis patients," *Decis. Support Syst.*, vol. 50, no. 2, 439–448, 2011.

[9] http://techemergence.com/deep-learning-medical-applications/

[10] P. P., et al "Conditional Deep Learning for energy-efficient and enhanced pattern recognition," in *2016 (DATE)*, 2016, pp. 475–480

[11] "PhysioBank Databases." [Online]. Available: https://www.physionet.org/physiobank/database/.