Distilling the Knowledge from Handcrafted Features for Human Activity Recognition

Zhenghua Chen, Le Zhang*, Zhiguang Cao, and Jing Guo

***Abstract*—Human activity recognition is a core problem in intelligent automation systems due to its far-reaching applications including ubiquitous computing, health-care services and smart living. Due to the non-intrusive property of smartphones, smartphone sensors are widely used for the identification of human activities. However, unlike applications in vision or data mining domain, feature embedding from deep neural networks performs much worse in terms of recognition accuracy than properly designed handcrafted features. In this paper, we posit that feature embedding from deep neural networks may convey complementary information and propose a novel knowledge distilling strategy to improve its performance. More specifically, an efficient shallow network, i.e. Single Layer Feedforward Neural Network (SLFN), with handcrafted features is utilized to assist a deep Long Short-Term Memory (LSTM) network. On the one hand, the deep LSTM network is able to learn features from raw sensory data to encode temporal dependencies. On the other hand, the deep LSTM network can also learn from SLFN to mimic how it generalize. Experimental results demonstrate the superiority of the proposed method in terms of recognition accuracy against several state-of-the-art methods in the literature.**

***Index Terms*—Human activity recognition, smartphone sensors, knowledge distilling, deep LSTM network, SLFN**

I. Introduction

Human activity recognition has attracted significant attention due to its importance in pervasive computing. It can benefit health-care services, for example monitoring the health conditions of elders [1], and providing exercise suggestions for users based on their daily activity levels [2]. Smart home research can make use of human activity information for energy efficient control [3]. Recently, the popular somatosensory game also requires the recognition of human activities [4].

A large number of sensors can be utilized for human activity recognition. Computer vision based systems have been studied for decades. A depth video based system for recognizing human activities was presented in [5]. Tran and Trivedi developed an automatic posture and gesture recognition system with an intelligent vision system [6]. However, vision based systems may encounter some problems, such as the impacts of illumination condition and privacy concerns. Radio frequency (RF) signals can also be used for the identification of human activities. Geng et al. proposed a human activity classification system using on-body RF characteristics [7]. Due to the complexity of RF signals, the activity recognition performance of RF based systems is limited. Moreover, wearable sensors are widely adopted for human activity recognition. Tao et al. presented a human activity recognition system with three-dimensional acceleration [8]. Gu et al. attempted to recognize both simple and complex activities using a wearable sensor node with accelerometer and environmental sensors for temperature, humidity and light level [9]. However, wearable sensors require extra costs for hardware and are intrusive for users. With the development of smartphone technology, modern smartphones are good sensing platforms with various sensors, such as accelerometer, gyroscope and barometer. Since they are widely available and non-intrusive for users, human activity recognition with smartphone sensors becomes more and more popular [10], [11].

Many algorithms can be leveraged to recognize human activities. The simple methodologies of Naive Bayes (NB) and K-Nearest Neighbors (KNN) have been applied to identify human activities with smartphone sensors [12]. More advanced machine learning algorithms of Artificial Neural Network (ANN) and Support Vector Machine (SVM) also achieved good performances on human activity recognition [13], [14]. All these shallow algorithms are trained on handcrafted features which may loss some important inherent information. In addition to the shallow models which rely on handcrafted features, a new branch of machine learning, i.e. deep learning [15], is able to learn representations from raw sensory data. The convolutional neural network was presented for activity recognition in [16], [17]. Another popular deep learning approach of Long Short-Term Memory (LSTM) was leveraged to recognition human activities using mobile devices in [18]. A combination of these two algorithms for human activity recognition can be found in [19]. However, the deep learning based approaches often suffer from noise effect and limited size of data, leading to an unsatisfactory performance.

With proper feature design, shallow models can achieve a good activity recognition performance. However, we can further improve it using deep neural network which contains complementary information by the use of feature embedding. Thus, in this paper, we propose to distill knowledge from handcrafted features with a deep LSTM network for human activity recognition using smartphone sensors. We empirically found that an efficient shallow network, i.e. Single Layer Feedforward Neural Network (SLFN), on top of properly designed handcrafted features outperforms deep neural networks.

Zhenghua Chen is with Department of Electrical and Electronic Engneering, Nanyang Technological University, Singapore, 50 Nanyang Ave, 639798 (e-mail: chen0832@e.ntu.edu.sg).

Le Zhang (corresponding author) is with Advanced Digital Sciences Center (ADSC), the Singapore-based research center of the University of Illinois at Urbana-Champaign (UIUC), 1 Fusionopolis Way, Sinagpore 138632 (e-mail: zhang.le@adsc.com.sg).

Zhiguang Cao and Jing Guo are with School of Automation, Guangdong University of Technology, 100 Waihuanxi Road, Higher Education Mega Center, Guangzhou, China 510006 (email: zhiguangcao@gdut.edu.cn, toguojing@gmail.com).

* indicates corresponding author.

However, deep neural networks still carry complementary information. Motivated by this, we propose a *knowledge distilling* strategy to employ SLFN with handcrafted features to assist supervising the deep LSTM network. Real experiments have been performed to show the effectiveness of the proposed approach. We also compare it with some state-of-the-art approaches for human activity recognition using smartphone sensors.

The main contributions of this paper are as follows:

- For the first time, we comprehensively compare the performance of shallow learning methods with handcrafted features and deep learning methods with raw data. These basic conclusions can serve as a general guidance for designing human activity analysis systems with smartphone sensors.
- We show that due to several factors such as the existence of noise and the scarceness of large dataset, deep LSTM performs worse than SLFN with proper feature engineering. We further design a knowledge distilling mechanism to improve the recognition accuracy of deep LSTM with the aid of SLFN with handcrafted features.
- We show that deep LSTM for feature embedding may carry complementary information. Hence, with the fusion of the deep LSTM and the shallow SLFN, we are able to outperform the state-of-the-art methods using real experiments.

The remaining of the paper is organized as follows: Section II presents the related works for human activity recognition with different algorithms. Section III illustrates feature engineering and deep recurrent neural network, followed by our proposed knowledge distillation for human activity recognition. Section IV first introduces the data for experiments and the experimental setup. Then, the experimental results and discussions are presented. Finally, Section V concludes this work.

II. Related Works

Human activity recognition can be divided into two categories based on different learning schemes. The first one is the shallow algorithms with handcrafted features and the second one is the deep learning algorithms with raw sensory data. We will review some related works for the two categories in the following paragraphs.

A. Shallow Algorithms

Eastwood and Jayne demonstrated some extensions of hyperbox neural network (HNN) which makes use of different modes of learning for human activity recognition [20]. The extension of combining unsupervised clustering and HNN achieved the best performance for human activity recognition. Wang et al. proposed a human activity recognition system using smartphone sensors with a feature selection algorithm which combines the filter and wrapper methods [12]. The simple algorithms of NB and KNN were employed to recognize six daily activities. They also investigated the human activity recognition performance with only acceleration or gyroscope and the combination of the two. Anguita et al. proposed a hardware friendly SVM (HF-SVM) based on fixed-point arithmetic for human activity classification [21]. The experiments showed that it is able to achieve a similar performance when compared to conventional SVM with less computational costs. Yang et al. presented an effective strategy for neural network classifier in recognizing human activities with acceleration data [13]. First, they applied a divide-and-conquer strategy to separate static and dynamics activities. Then, a common principle component analysis (PCA) approach was employed to reduce the dimension of the feature sets. Finally, a neural network model was applied with the feature subsets to recognize the static or dynamic activities. Ronao and Cho proposed to recognize human activities using a two-stage continuous hidden Markov model (CHMM) where the two levels of CHMM were used for coarse and fine classifications, respectively [22]. Rana et al. improved the sparse random classifier with singular value decomposition (SRC-SVD) to identify human activities [23]. The SVD was utilized to construct the random projection matrix for SRC. Real experiments demonstrated a superior performance over the conventional SRC. Seera et al. proposed a hybrid of fuzzy min-max (FMM) neural network and the classification and regression tree (CART) for the identification of human activities [24]. In their proposed system, the FMM was mainly used for data incremental learning and the CART was applied to provide explanations for the predictions. Chen et al. proposed a human activity recognition system regardless of device orientations based on coordinate transformation and PCA [11]. After pre-processing with their proposed scheme dealing with orientation variations, four generic classifiers, i.e. Decision Tree (DT), KNN, ANN and SVM, were leveraged to identify five common daily activities. They also presented an efficient online SVM algorithm in handling the inherent difference of signals for different placements and subjects.

B. Deep Learning

Li et al. proposed a channel-wise feature extraction with Sparse Auto-Encoders (SAEs) denoted as SAEs-c for human activity recognition [25]. Here, the three axes of acceleration/gyroscope and the magnitudes of the three-dimensional acceleration/gyroscope are viewed as different channels. Ronao and Cho proposed a Convolutional Neural Network (Convnet) based human activity recognition system with smartphone sensors [17]. The convolutional operation can automatically extract robust features for activity classification. They further improve the performance of their system by using additional information from fast Fourier transform of the raw data. Moreover, they also explored the use of handcrafted features with Convnet to improve the recognition performance of human activities [26]. Tao et al. presented a Bi-directional Long Short-Term Memory (BLSTM) network with two-directional features for human activity recognition [18]. To improve the performance of their proposed system, they leveraged on an ensemble scheme of BLSTM which combines different properties of acceleration signals. Ordóñez and Roggen presented a deep convolutional LSTM network for human activity recognition [19]. The deep convolutional operation can extract representative features from raw sensory data.

And the LSTM network can encode temporal dependencies to the features extracted by the deep convolutional network for the final identification of human activities. They also showed that their proposed framework can be used for different sensor modalities individually and the combination of them.

The shallow models with a proper feature engineering are able to perform well for human activity recognition. However, handcrafted features designed with domain knowledge would inevitably miss some implicit key features. The deep learning approaches can distill knowledge from handcrafted features to improve its performance in identifying human activities. In addition, they still carry complementary information for human activity classification. Thus, the fusion of deep learning approaches with the shallow models will boost the performance of human activity recognition.

III. Methodology

A. Notations

A word about our notation and background material. All vectors will be column vectors unless transposed to a row vector by a prime superscript $\top$. Data samples are presented in a M dimensional space: $\mathbf{x}_i \in \mathcal{X} \subseteq I\!R^M, i \in \{1, \ldots, N\}$. Let $\mathbf{X} \in I\!R^{N \times M}$ be an $N \times M$ matrix obtained by stacking N samples in $\mathcal{X}$, *i.e.*, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top$. Let $\mathbf{Y} \in \{y_1, y_2, ..., y_N\}^\top$ be a vector obtained by stacking the labels of samples in $\mathcal{X}$, where y_i takes a value from the set of class labels $\{\omega_1, \omega_2, ..., \omega_C\}$. We consider C classes classification problem. We will denote a generic data point as $\mathbf{x}$ by ignoring the subscript for convenience, and use $\mathbf{x}_\diamond$ with $\diamond$ denoting the placeholder for the index wherever necessary. We use $\mathbf{x}_{\diamond\bullet}$ to indicate the feature subset indexed by $\bullet$ of the $\diamond^{th}$ data.

B. Feature Engineering

The time series of smartphone sensor data cannot be directly used for shallow models to identify different human activities. Researches attempt to leverage on sliding windows with overlapping to obtain segments of raw sensory data where representative features can be extracted [12]. This process is known as feature engineering which is vital for the success of using generic shallow algorithms [27]. The final objective of feature engineering is to generate representations which are able to capture slight differences among different activities so that they can be separated using shallow algorithms, such as KNN, ANN and SVM.

In this work, we employ statistical handcrafted features from time and frequency domains to capture the properties of different activities in the two domains. For instance, the feature of mean can reflect the magnitude of signals, which is helpful in separating static activities from dynamic ones. The features in frequency domains can be leveraged to separate activities with different frequencies, such as walking and running. With domain knowledge, some representative features can be designed. Then, the shallow algorithms with these well designed features have been shown to be effective for human activity recognition [21], [22], [24], [23]. The detailed handcrafted features which have been widely used [28], [20], [12] can be found in TABLE I.

TABLE I
Handcrafted Features

Domain	Features
Time	Mean value
	Standard deviation
	Maximum
	Minimum
	Signal magnitude area
	Average sum of the squares
	Interquartile range
	Signal Entropy
	Autorregresion coefficients
	Correlation coefficient
Frequency	Largest frequency component
	Weighted average
	Skewness
	Kurtosis
	Energy of a frequency interval
	Angle between two vectors

C. Deep Recurrent Neural Network

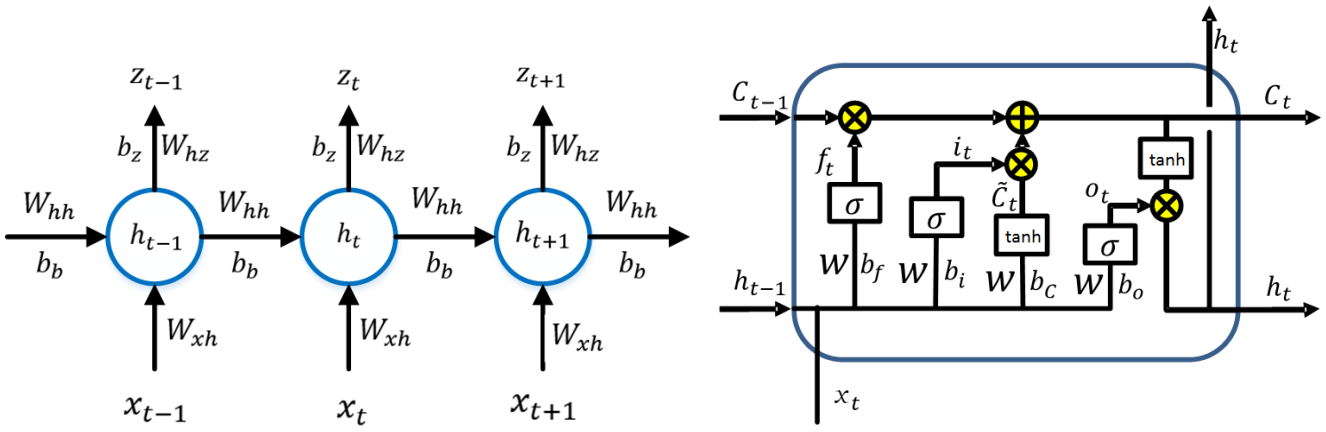
Signals from smartphone sensors turn out to be highly noisy, generating large intra-class variance for human activity recognition. Learning a discriminative feature embedding in a data-driven fashion can partly alleviate this problem. Learning based representation, which is usually rebranded as deep learning recently, makes it possible for a machine to automatically discover the representations needed for a given task on top of raw data. Human activities can be hierarchical because complex activities are composed of temporal involvement of simple actions. For example, "Walking" can be modeled by periodic pattern where the body vaults over the stiff limb or limbs with each step. In this study, we introduce a recurrent structure which can be "doubly deep" in that it learns compositional representations in both space and time. Learning long-term dependencies which is beneficial for human activity recognition can be effectively encapsulated into our framework when non-linearities are incorporated into the network state updates.

Traditional recurrent neural network (RNN) in Fig. 1(a) models temporal dynamics by mapping input sequences to hidden states, and hidden states to outputs via the following recurrence equations:

$$\begin{aligned} \mathbf{h}_t &= g(\mathbf{W}_{\mathbf{x}h}\mathbf{x}_t + \mathbf{W}_{hh}\mathbf{h}_{t-1} + b_b) \\ z_t &= g(\mathbf{W}_{hz}\mathbf{h}_t + b_z) \end{aligned} \tag{1}$$

where g stands for non-linear activation function (such as *ReLU* or *hyperbolic tangent*), $\mathbf{x}_t$ and $\mathbf{h}_t$ represent the input signal and hidden states, respectively, z_t is the output at time t, and $\{\mathbf{W}_{\mathbf{x}h}, \mathbf{W}_{hh}, \mathbf{W}_{hz}\}$ and $\{b_b, b_z\}$ are weights and biases, respectively.

The RNN has been reported to be efficient on various topics such as speech recognition [29] and text generation [30]. However, they turn out to be difficult to optimize because of the well-known issue of vanishing and exploding gradients [31]. The *Long Short-Term Memory* (LSTM) approach proposed in [31] remedies these issues based on memory units where learning of when to "forget" and/or "update" previous hidden states is made possible.As illustrated in Fig. 1(b), LSTM



(a) Conventional RNN (b) LSTM

Fig. 1. Structure of RNN and LSTM.

updates itself at time t based on its input $\mathbf{x}_t$, $\mathbf{h}_{t-1}$, and $\mathbf{C}_{t-1}$ by way of:

$$\begin{aligned}
\mathbf{i}_t &= \sigma(\mathbf{W}_{xi}\mathbf{x}_t + \mathbf{W}_{hi}\mathbf{h}_{t-1} + b_i) \\
\mathbf{f}_t &= \sigma(\mathbf{W}_{xf}\mathbf{x}_t + \mathbf{W}_{hf}\mathbf{h}_{t-1} + b_f) \\
\mathbf{o}_t &= \sigma(\mathbf{W}_{xo}\mathbf{x}_t + \mathbf{W}_{ho}\mathbf{h}_{t-1} + b_o) \\
\tilde{\mathbf{C}}_t &= tanh(\mathbf{W}_{xc}\mathbf{x}_t + \mathbf{W}_{hc}\mathbf{h}_{t-1} + b_C) \\
\mathbf{C}_t &= \mathbf{f}_t \bigodot \mathbf{C}_{t-1} + \mathbf{i}_t \bigodot \tilde{\mathbf{C}}_t \\
\mathbf{h}_t &= \mathbf{o}_t \bigodot tanh(\mathbf{C}_t)
\end{aligned} \tag{2}$$

where $\sigma(\mathbf{x}) = (1 + e^{-\mathbf{x}})^{-1}$ and $tanh(\mathbf{x}) = \frac{e^{\mathbf{x}} - e^{-\mathbf{x}}}{e^{\mathbf{x}} + e^{-\mathbf{x}}} = 2\sigma(2\mathbf{x} - 1)$ stand for *sigmoid* and *hyperbolic tangent* non-linearities, respectively. The former squashes real-valued inputs to a $[0, 1]$ range while the latter similarly squashing its inputs to $[1, 1]$. $\bigodot$ means element-wise product. $\mathbf{W}$ and b are the weights and bias, respectively. $\{\mathbf{h}_t, \mathbf{i}_t, \mathbf{f}_t, \mathbf{o}_t, \tilde{\mathbf{C}}_t, \mathbf{C}_t\}$ represent hidden units, input gate, forget gate, output gate, input modulation gate and memory gate, respectively. The memory cell unit $\mathbf{C}_t$ is consists of two component. One is the previous memory cell unit $\mathbf{C}_{t-1}$ modulated by $\mathbf{f}_t$. The other is $\tilde{\mathbf{C}}_t$ which is modeled by the current input and previous hidden state, modulated by the input gate $\mathbf{i}_t$. The sigmoidal nature of $\mathbf{i}_t$ and $\mathbf{f}_t$ squashes themselves into a range of $[0, 1]$. They can be regarded as knobs that LSTM learns to selectively forget its previous memory or consider its current input. In the same way, the output gate $\mathbf{o}_t$ models the transfer from memory cells to hidden states. Based on these mechanisms, LSTM learns complex as well as long-term temporal dynamics that exists in raw smartphone sensor data. In our implementation, we employ a deeper structure by stacking two sets of LSTM to further enhance the discriminative ability of the feature embedding.

D. Knowledge Distillation

1) Rationale: Learning representations can significantly outperform commonly used machine learning approaches such as RF, SVM, etc. However, we empirically found that the performance of deep LSTM was still far from satisfactory. We hypothesis that there are two reasons. The first one, as we mentioned above, signals from smartphone sensors turn out to be highly noisy, generating large intra-class variance. The other reason is the lacking of very large-scale datasets. It is widely accepted that deep neural network with a large number of parameters relies on a large number of training samples to optimize. We empirically found that with proper feature design, shallow SLFN can achieve even better recognition accuracy. However, we argue that feature embedding from deep LSTM can represent complementary information which can be beneficial in our task. This may be because of inevitable information lost in handcrafted features. In order to further improve the recognition performance, we propose to distill knowledge from handcrafted features.

Motivated by [32], we propose to use the class probabilities produced by the shallow model of SLFN on top of handcrafted features as "soft targets" for training the deep model of LSTM. In this way, one may benefit from rich information provided by the high-entropy output of SLFN. Generally speaking, the training process of classification based neural network is to maximize the average log probability of correct answers. As a side-effect of the training process, the resulting neural network assigns probabilities to all of incorrect answers. In most cases, the probabilities of the incorrect answers should be much less than the ones from the correct answers. This also indicates how the shallow model on top of handcrafted features tends to generalize. For example, an instance of "Walking" may have a higher chance to be misclassified as "Walking Upstairs" or "Walking Downstairs", however, it can almost not be predicted as an activity of "Laying". In this case, the neural network may provide much more information per training case than hard targets.

In order to generalize well on unseen testing data, neural networks are usually trained to minimize the loss function on a similar set of training data. We prefer models which generalize well. However, it is usually unknown about the correct way to generalize. When distilling the knowledge from another model, which we named it as "teacher model", into a much deeper one, we can somehow enforce the deep model to generalize in the same way as the teacher model. If the teacher

model generalizes well because, for example, it carries rich domain knowledge which is invisible from raw data directly, our deeper model trained on raw data may generalize in the same way. In this manner, it is likely that the deeper model may lead to better performance on test data than the model trained directly on the same raw data.

2) Proposed method: Neural networks typically produce class probabilities by using "softmax" operation on the output neurons:

$$\begin{aligned}\mathbf{p}_i(\mathbf{y}|\mathbf{x};\theta) =& \mathbf{p}(\mathbf{y}_i|\mathbf{x};\theta)\\ =& \textbf{softmax}(\mathbf{g}(\mathbf{x};\theta))\\ =& \frac{exp(\mathbf{Z}_i/T)}{\sum_j exp(\mathbf{Z}_j/T)},\end{aligned} \tag{3}$$

where the logit $\mathbf{Z}_j$ stands for the output value of the j^{th} output neuron and $\mathbf{p}_i$ represents the probability of the i^{th} class. Higher value for T, which means the temperature and is usually set to be 1, leads to a smoother probability for each class.

The parameters θ are learned by minimizing the log loss for all training samples by the following manner:

$$\begin{aligned}L^{hard} =& \sum_{n=1}^{N} L(\mathbf{x}_n, \mathbf{y}_n;\theta)\\ =& \sum_{n=1}^{N}\sum_{i=1}^{C} \mathbf{1}\{\mathbf{y}_n = c\}\textbf{log } \mathbf{p}(\mathbf{y}_i|\mathbf{x};\theta)\end{aligned} \tag{4}$$

Our idea is to train the deep LSTM model using the resulting distribution in Eqn. (4) rather than the ground truth labels. We can simply increase the entropy of the posteriors in Eqn. (3) by raising the temperature. More formally,

$$\mathbf{y}_i^{target} = \mathbf{g}(\mathbf{x};\theta)_i^{student} = \frac{(\mathbf{y}_i^{teacher})^{\frac{1}{T}}}{\sum_{k=1}^{C}(\mathbf{y}_k^{teacher})^{\frac{1}{T}}} \tag{5}$$

where superscript $student$ and $teacher$ denote the student model (deep LSTM trained with raw data) and the teacher model (SLFN trained with handcrafted features), respectively. T is a temperature parameter mentioned above. Now the new loss function for the deep LSTM model can be set as:

$$L^{soft} = \sum_{n=1}^{N} L(\mathbf{x}_n, \mathbf{y}_n;\theta^s) = -\sum_{n=1}^{N}\sum_{i=1}^{C} \mathbf{y}_{n,c}^{target}\textbf{log } \mathbf{p}(\mathbf{y}_i|\mathbf{x};\theta^s) \tag{6}$$

Motivated by [32], we also found that a much better way is to use a weighted average of two different objective functions:

$$L = \alpha L^{soft} + (1-\alpha)L^{hard} \tag{7}$$

3) An overview of the proposed system: Although deep LSTM can be improved by the improved objective function in Eqn. (7), shallow neural network may still convey complementary information. In order to fully exploit this, we propose to fuse both deep LSTM and shallow SLFN. More specifically, as illustrated in Fig. 2, our system predict the category of an instance as:

$$\mathbf{y} = j = \text{argmax } (\mathbf{g}^s(\mathbf{x}^s,\theta^s) + \mathbf{g}^t(\mathbf{x}^t,\theta^s)) \tag{8}$$

The teacher net is an efficient shallow network, i.e. SLFN, working on handcrafted features $\mathbf{x}^t$. A *ReLU* activation function is utilized instead of the commonly used *sigmoid* activation function. It has 100 neurons for the hidden layer. We use dropout [33] with a ratio of 0.5 in the hidden layer to reduce overfitting. The Stochastic Gradient Descent (SGD) with an initial learning rate of e^{-4} and e^{-5} after $40K$ iterations is applied for parameter optimization. The maximum iteration is set to be $80K$. The student net is realized by a deep recurrent network which stacks two LSTM layers. In each layer, there are 128 LSTM modules to model the temporal involvements of different activities. The technique of RMSprop is employed to optimize the whole parameter with a batchsize of 200. We set the learning rate to be e^{-4}. Motivated by deeply supervised network [34], we put two separated fully connected layers followed by a softmax layer on each LSTM layer. In this way, the error signal can be easily propagated to the lower layer of LSTM to reduce gradient vanishing. Finally, we use a linear combination of the three outputs to give a final classification.

4) Differences with [32]: Our proposed method bears some similarity with [32] but differentiates from it significantly. Firstly, the method in [32] addressed the problem of model compression where a shallow model was trained to imitate an ensemble of deep models. Here, our method is based on an observation that shallow models trained on handcrafted features usually tend to perform better but still convey complementary information compared with deep models. Our method works in a synergistic manner where deep models can be enhanced. Secondly, the method in [32] works in a single-view scenario where both teacher and student models embark on the same feature set. However, our method works in a multi-view setting where teacher and student models work on handcrafted features and raw data, respectively.

IV. Evaluation

In this section, we first introduce the data for experiments and the experimental setup. Then, the experimental results are presented and discussed.

A. Data Acquisition

The smartphone sensor data was collected from 30 volunteers with ages from 19 to 48. A Samsung Galaxy SII smartphone attached to the waist of subjects is utilized for experiments. The performed activities include *standing*, *sitting*, *laying*, *walking*, *walking downstairs* and *walking upstairs*, which are the most common daily activities in life. A protocol was designed for data collection. Each subject followed the protocol twice. In the first round, the subject was asked to fix the smartphone on the left side of the belt; and in the second round, the smartphone can be put by the subject as preferred. During adjacent tasks, the subject was told to rest for 5 seconds. The ground truth activities are generated from visual interface. The data collection was performed in laboratory conditions, but the subjects were asked to perform freely according to the protocol for a more naturalistic dataset. TABLE II presents the experiment protocol [28].

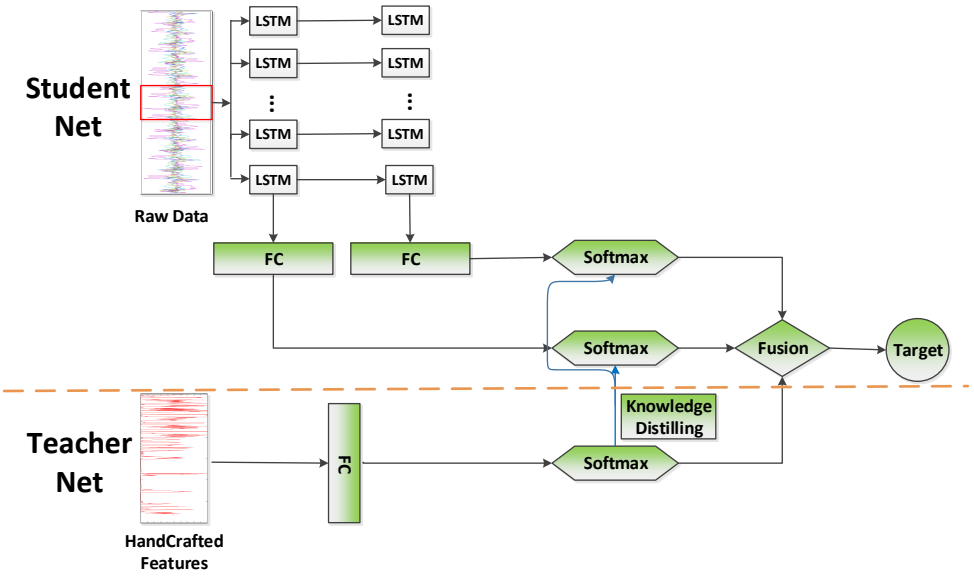


Fig. 2. An overview of the proposed system. We propose to use handcrafted feature based SLFN (teacher net) to further supervise raw data based deep LSTM (student net).

TABLE II
THE EXPERIMENT PROTOCOL FOR DATA COLLECTION.

No.	Static	Time (sec)	No.	Dynamic	Time (sec)
0	Start (Standing Pos)	0	7	Walk (1)	15
1	Stand (1)	15	8	Walk (2)	15
2	Sit (1)	15	9	Walk Downstairs (1)	12
3	Stand (2)	15	10	Walk Upstairs (1)	12
4	Lay Down (1)	15	11	Walk Downstairs (2)	12
5	Sit (2)	15	12	Walk Upstairs (2)	12
6	Lay Down (2)	15	13	Walk Downstairs (3)	12
			14	Walk Upstairs (3)	12
			15	Stop	0
				Total	**192**

The data including three-dimensional linear acceleration, total acceleration and gyroscope was collocated with a sampling rate of $50Hz$. We perform a sliding window of 2.56 seconds with an overlapping of 50%, which means each segment contains 128 data samples with 9 dimensions (linear acceleration, total acceleration and gyroscope). The handcrafted features in TABLE I are extracted for each segment. Totally, we have 10299 samples where 70% of them are selected for training and the remaining for testing. A more detailed description of the data can be found in [28].

B. Experimental Setup

To verify the effectiveness of the proposed approach, we have compared it with several advanced methods including the shallow machine learning algorithms of SLFN, random vector functional link network (RVFL) [35], RF [36] and SVM [37]. We also compare our proposed method with several state-of-the-art approaches in the literature. Several ablation studies are also carried out for a better understanding of the propose method. Parameters for different methods are tuned based on a validation set which is randomly selected from training set. After parameter tuning, we re-train each method on the whole training set and evaluation it on the test set.

C. Experimental Results

1) Raw data versus handcrafted features for shallow algorithms: In order to demonstrate the superiority of the handcrafted features over the raw data, we empirically compare several state-of-the-art shallow learning methods in TABLE III. More specifically, we consider SVM [37], RF [36], RVFL [35] and SLFN. From TABLE III, it is not surprising that all the methodologies with handcrafted features outperform these with raw data. Although information loss is inevitable, handcrafted feature extraction can largely reduce noise and overfitting hence making the learned classifiers more generalizable. The efficient SLFN presents a superior performance over the other learning algorithms with both handcrafted features and raw data.

TABLE III
RAW DATA VERSUS HANDCRAFTED FEATURES FOR SHALLOW ALGORITHMS

Method	Raw Data	Handcrafted Features
SVM [37]	76.7%	94.0%
RF [36]	85.0%	93.0%
RVFL [35]	73.3%	94.9%
SLFN	91.2%	96.5%

2) Raw data versus handcrafted features for deep methods: In this part, we compare the performance of different input modalities under the context of deep learning. What we consider are multi-layer feedforward deep neural networks (DNN). More specifically, we use the widely used *ReLU* as activation function followed by multi-way softmax as classifier. Please note that deep LSTM is not applicable here as temporal information is totally lost in handcrafted features. We present the results in TABLE IV. *DNN-4* means a 4 layer deep feedforward network which contains 2 hidden layers. We can observe that DNNs with raw data can achieve slightly better performance than these with handcrafted features. But the results are far from satisfactory when compared with shallow algorithms with handcrafted features. However, we argue that deep models still contain some complementary information by the use of feature embedding.

TABLE IV
RAW DATA VERSUS HANDCRAFTED FEATURES FOR DEEP METHODS

Method	Raw Data	Handcrafted Features
LSTM	91.5%	N.A.
DNN-4	90.1%	88.5%
DNN-5	90.9%	90.2%

3) Number of training samples for deep methods: Since deep models suffer from the scarceness of large training data, in order to demonstrate this, we evaluate the performance of deep LSTM and the proposed approach with different number of training samples. The results are shown in TABLE V. More specifically, randomized stratified sampling is employed to generate a training subset with the same distribution for each class. Experimental results indicate that deep neural networks tend to perform better with more training data. Owing to the proposed knowledge distilling from handcrafted features and the fusion framework, the performance of the proposed method is not very sensitive to the number of training samples.

TABLE V
PERFORMANCE OF DEEP LSTM AND THE PROPOSED METHOD WITH DIFFERENT NUMBER OF TRAINING SAMPLES

Number of Training Instances	Deep LSTM	Proposed Method
1000	76.4%	96.6%
2000	85.2%	97.2%
3000	89.0%	97.4%
4000	89.5%	97.5%
5000	90.1%	97.5%
6000	90.9%	97.6%
7352 (all training samples)	91.5%	97.7%

4) Results of the proposed method: Here, we present the performance of our proposed method and compare it to several state-of-the-art methods in the literature. The state-of-the-art methods can be divided into two categories, i.e. shallow algorithms and deep algorithms. The shallow models include HNN [20], FW KNN [12], FW Naive Bayes [12], HF-SVM [21], Two-stage CHMM [22], SRC-SVD [23], and FMM-CART [24]; and the deep models contain SAEs-c [25], Convnet [17], HCF Convnet [26] and tFFT Convnet [17]. The detailed reviews of all these approaches can be found in Section II.

The results of the proposed method and the state-of-the-art methods are shown in TABLE VI. Since our proposed approach takes the merits of both shallow algorithms with handcrafted features and deep algorithms with feature embedding, it outperforms all the state-of-the-art algorithms in the literature. This indicates the effectiveness of the proposed approach that combines both shallow and deep algorithms with knowledge distilling mechanism.

Since the proposed approach is a deep learning based approach, the time complexity will be a big concern. Therefore, we tested the training and testing time of the proposed approach. Note that the shallow algorithms shall work much faster than the proposed deep learning based approach. Our code is running on a server with a GPU of GeForce GTX TITAN. The training time is 49 minutes 23 seconds. However, this tedious training process only needs to be done once. The testing time for all the testing samples (2947) is 0.42 seconds. It means that the testing time for each sample is $0.42/2947$ seconds which can be neglected. Thus, we can claim that our proposed approach can achieve real-time human activity recognition.

TABLE VI
COMPARISON OF THE PROPOSED METHOD WITH THE STATE-OF-THE-ART APPROACHES.

Method		Accuracy
Shallow	HNN [20]	87.4%
	FW KNN [12]	87.8%
	FW Naive Bayes [12]	90.1%
	HF-SVM [21]	89%
	Two-stage CHMM [22]	91.76%
	SRC-SVD [23]	95%
	FMM-CART [24]	96.52%
Deep	SAEs-c [25]	92.16%
	Convnet [17]	94.79%
	HCF Convnet [26]	95.75%
	tFFT Convnet [17]	95.75%
	Proposed Method	97.7%

5) Ablation studies of the proposed method: We present several ablation studies to investigate the impacts of different components in our system. Firstly, we show that although deep models perform slightly worse than shallow models on handcrafted features, it conveys complementary information which can be better utilized to boost the final performance. Then, we indicate that, with the proposed knowledge distilling mechanism, we are able to improve the performance of deep LSTM for human activity recognition. Finally, we investigate the performance of alternative architecture variants of the proposed method.

Complementary Information From Different Models: Firstly, we observe that deep LSTM and shallow SLFN lead

to an accuracy of 91.5% (before knowledge distilling) and 96.5%, respectively. Due to several reasons aforementioned, deep LSTM performs worse than SLFN with handcrafted features for human activity analysis. We illustrated the confusion matrix of these two approaches in Fig. 3 and Fig. 4. It can be found that deep LSTM performs better at distinguishing *walking* and *walking upstairs*. On the other hand, shallow SLFN turns out to be more accurate in differentiating *sitting* and *standing*. Moreover, a better accuracy of 97.7% is achieved by our proposed approach which fuses these two models. The confusion matrix of the proposed approach is shown in Fig. 5. We can find that the proposed approach can better distinguish all the activities.

Effect of Knowledge Distilling: We observe that deep LSTM leads to an accuracy of 91.5% (see TABLE IV) by training from scratch using raw data. When trained with knowledge distilling, deep LSTM (without fusion) is able to achieve 92.9%. This clearly demonstrate the superiority of the proposed knowledge distilling, as mentioned in Section III-D.

Role of Teacher and Student Net: It is natural and beneficial to use a model with better accuracy as teacher net. To prove this, we also conduct another experiment with the variant, which works the other way around, by using deep network as teacher net and shallow net as student net. This leads to a slightly worse overall recognition accuracy of 97.3%.

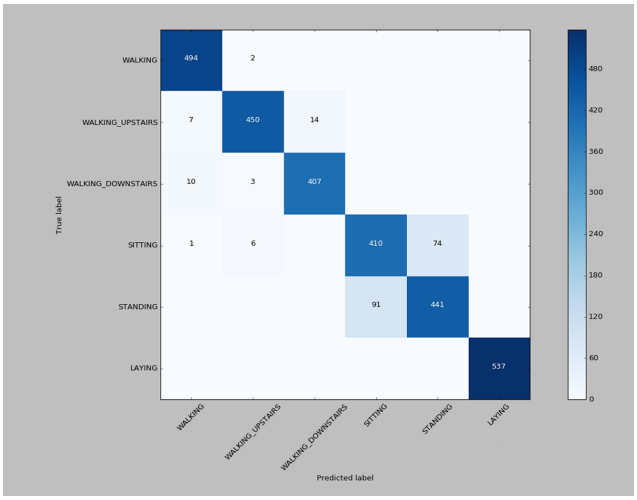


Fig. 3. Confusion Matrix of deep LSTM network.

V. Conclusion

This paper adds value to the literature of human activity recognition with smartphone sensors in a significant manner. Firstly, we conduct comprehensive comparisons for handcrafted features and raw data on both shallow and deep models. The shallow models with handcrafted features have a superior performance over these with raw data, which clearly indicates the importance of handcrafted features for shallow models. For deep models, the handcrafted features slightly degrades its performance. These basic conclusions can serve as a general guidance for the design of human activity recognition systems

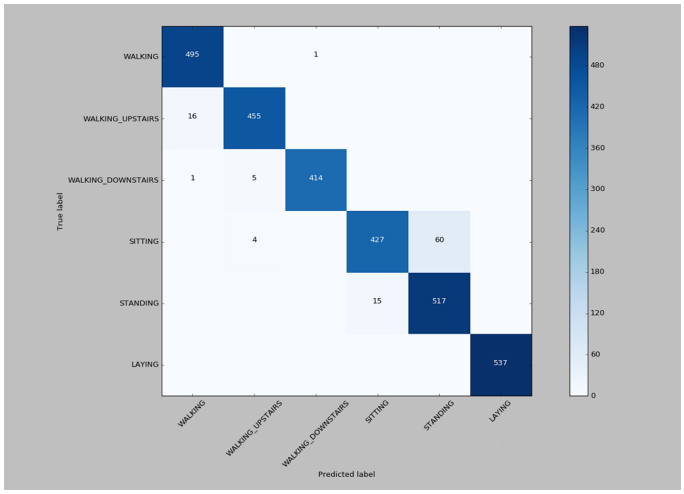


Fig. 4. Confusion Matrix of SLFN.

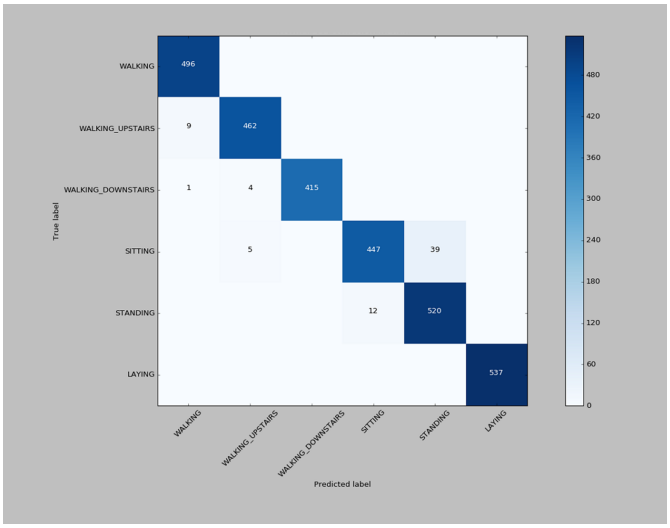


Fig. 5. Confusion Matrix of the proposed method

using smartphone sensors. Due to several factors such as noise and the scarceness of large dataset, the results of deep models with raw data are far from satisfactory when compared with shallow models with proper feature design. Thus, we further propose a knowledge distilling mechanism to improve the recognition accuracy of a deep network, i.e. deep Long Short-Term Memory (LSTM), with the aid of an efficient shallow model, i.e. Single Layer Feedforward Neural Network (SLFN). The recognition performance has been significantly improved with the proposed knowledge distilling. Finally, since the deep LSTM contains complementary information, we show that with the fusion of deep LSTM and SLFN, we are able to outperform the state-of-the-art methods in the literature. The final human activity recognition accuracy of the proposed method is as high as 97.7% with smartphone sensors.

References

[1] H. Ghasemzadeh and R. Jafari, "Physical movement monitoring using body sensor networks: A phonological approach to construct spatial decision trees," *IEEE Transactions on Industrial Informatics*, vol. 7, no. 1, pp. 66–77, 2011.

[2] N. Alshurafa, W. Xu, J. J. Liu, M.-C. Huang, B. Mortazavi, C. K. Roberts, and M. Sarrafzadeh, "Designing a robust activity recognition framework for health and exergaming using wearable sensors," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 5, pp. 1636–1646, 2014.

[3] A. Javed, H. Larijani, A. Ahmadinia, and D. Gibson, "Smart random neural network controller for hvac using cloud computing technology," *IEEE Transactions on Industrial Informatics*, vol. 13, no. 1, pp. 351–360, 2017.

[4] A. Tachibana, J. A. Noah, S. Bronner, Y. Ono, and M. Onozuka, "Parietal and temporal activity during a multimodal dance video game: an fnirs study," *Neuroscience Letters*, vol. 503, no. 2, pp. 125–130, 2011.

[5] A. Jalal, M. Z. Uddin, and T.-S. Kim, "Depth video-based human activity recognition system using translation and scaling invariant features for life logging at smart home," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 3, 2012.

[6] C. Tran and M. M. Trivedi, "3-d posture and gesture recognition for interactivity in smart spaces," *IEEE Transactions on Industrial Informatics*, vol. 8, no. 1, pp. 178–187, 2012.

[7] Y. Geng, J. Chen, R. Fu, G. Bao, and K. Pahlavan, "Enlighten wearable physiological monitoring systems: On-body rf characteristics based human motion classification using a support vector machine," *IEEE transactions on mobile computing*, vol. 15, no. 3, pp. 656–671, 2016.

[8] D. Tao, L. Jin, Y. Wang, and X. Li, "Rank preserving discriminant analysis for human behavior recognition on wireless sensor networks," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 1, pp. 813–823, 2014.

[9] T. Gu, L. Wang, Z. Wu, X. Tao, and J. Lu, "A pattern mining approach to sensor-based human activity recognition," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 9, pp. 1359–1372, 2011.

[10] J.-H. Hong, J. Ramos, and A. K. Dey, "Toward personalized activity recognition systems with a semipopulation approach," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 101–112, 2016.

[11] Z. Chen, Q. Zhu, C. S. Yeng, and L. Zhang, "Robust human activity recognition using smartphone sensors via ct-pca and online svm," *IEEE Transactions on Industrial Informatics*, 2017.

[12] A. Wang, G. Chen, J. Yang, S. Zhao, and C.-Y. Chang, "A comparative study on human activity recognition using inertial sensors in a smartphone," *IEEE Sensors Journal*, vol. 16, no. 11, pp. 4566–4578, 2016.

[13] J.-Y. Yang, J.-S. Wang, and Y.-P. Chen, "Using acceleration measurements for activity recognition: An effective learning algorithm for constructing neural classifiers," *Pattern Recognition Letters*, vol. 29, no. 16, pp. 2213–2220, 2008.

[14] A. Fleury, M. Vacher, and N. Noury, "Svm-based multimodal classification of activities of daily living in health smart homes: sensors, algorithms, and first experimental results," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 2, pp. 274–283, 2010.

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

[16] W. Jiang and Z. Yin, "Human activity recognition using wearable sensors by deep convolutional neural networks," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1307–1310.

[17] C. A. Ronao and S.-B. Cho, "Human activity recognition with smartphone sensors using deep learning neural networks," *Expert Systems with Applications*, vol. 59, pp. 235–244, 2016.

[18] D. Tao, Y. Wen, and R. Hong, "Multi-column bi-directional long short-term memory for mobile devices-based human activity recognition," *IEEE Internet of Things Journal*, 2016.

[19] F. J. Ordóñez and D. Roggen, "Deep convolutional and lstm recurrent neural networks for multimodal wearable activity recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[20] M. Eastwood and C. Jayne, "Evaluation of hyperbox neural network learning for classification," *Neurocomputing*, vol. 133, pp. 249–257, 2014.

[21] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *International Workshop on Ambient Assisted Living*. Springer, 2012, pp. 216–223.

[22] C. A. Ronao and S.-B. Cho, "Human activity recognition using smartphone sensors with two-stage continuous hidden markov models," in *Natural Computation (ICNC), 2014 10th International Conference on*. IEEE, 2014, pp. 681–686.

[23] R. Rana, B. Kusy, J. Wall, and W. Hu, "Novel activity classification and occupancy estimation methods for intelligent hvac (heating, ventilation and air conditioning) systems," *Energy*, vol. 93, pp. 245–255, 2015.

[24] M. Seera, C. K. Loo, and C. P. Lim, "A hybrid fmm-cart model for human activity recognition," in *2014 IEEE International Conference on Systems, Man and Cybernetics*. IEEE, 2014, pp. 182–187.

[25] Y. Li, D. Shi, B. Ding, and D. Liu, "Unsupervised feature learning for human activity recognition using smartphone sensors," in *Mining Intelligence and Knowledge Exploration*. Springer, 2014, pp. 99–107.

[26] C. A. Ronao and S.-B. Cho, "Deep convolutional neural networks for human activity recognition with smartphone sensors," in *International Conference on Neural Information Processing*. Springer, 2015, pp. 46–53.

[27] H. Liu and H. Motoda, *Feature extraction, construction and selection: A data mining perspective*. Springer Science & Business Media, 1998.

[28] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones." in *ESANN*, 2013.

[29] X. Chen, X. Liu, Y. Wang, M. J. Gales, and P. C. Woodland, "Efficient training and evaluation of recurrent neural network language models for automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 11, pp. 2146–2157, 2016.

[30] I. Sutskever, J. Martens, and G. E. Hinton, "Generating text with recurrent neural networks," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, 2011, pp. 1017–1024.

[31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[32] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[33] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[34] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Artificial Intelligence and Statistics*, 2015, pp. 562–570.

[35] L. Zhang and P. N. Suganthan, "A comprehensive evaluation of random vector functional link networks," *Information Sciences*, vol. 367, pp. 1094–1105, 2016.

[36] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[37] A. J. Smola and B. Schölkopf, *Learning with Kernels*. GMD-Forschungszentrum Informationstechnik, 1998.