

Received October 22, 2017, accepted November 23, 2017, date of publication December 4, 2017, date of current version February 14, 2018.

Digital Object Identifier 10.1109/ACCESS.2017.2779939

LSTM Fully Convolutional Networks for Time Series Classification

FAZLE KARIM¹, SOMSHUBRA MAJUMDAR², HOUSHANG DARABI¹, (Senior Member, IEEE), AND SHUN CHEN¹

¹Mechanical and Industrial Engineering, University of Illinois at Chicago, Chicago, IL 60607, USA

²Computer Science, University of Illinois at Chicago, Chicago, IL 60607, USA

Corresponding author: Houshang Darabi (hdarabi@uic.edu)

ABSTRACT Fully convolutional neural networks (FCNs) have been shown to achieve the state-of-the-art performance on the task of classifying time series sequences. We propose the augmentation of fully convolutional networks with long short term memory recurrent neural network (LSTM RNN) sub-modules for time series classification. Our proposed models significantly enhance the performance of fully convolutional networks with a nominal increase in model size and require minimal preprocessing of the data set. The proposed long short term memory fully convolutional network (LSTM-FCN) achieves the state-of-the-art performance compared with others. We also explore the usage of attention mechanism to improve time series classification with the attention long short term memory fully convolutional network (ALSTM-FCN). The attention mechanism allows one to visualize the decision process of the LSTM cell. Furthermore, we propose refinement as a method to enhance the performance of trained models. An overall analysis of the performance of our model is provided and compared with other techniques.

INDEX TERMS Convolutional neural network, long short term memory recurrent neural network, time series classification.

I. INTRODUCTION

Over the past decade, there has been an increased interest in time series classification. Time series data is ubiquitous [1], existing in weather readings [2], financial recordings [3], industrial observations [4], and psychological signals [5], [6]. Several approaches, including feature-based [7], ensembles [8]–[10], and deep learning [11], [12], have been utilized to classify time series. Deep learning has been successfully utilized in various applications that require time series data, especially in control systems [13], [14]. In this paper, two deep learning models to classify time series datasets are proposed, both of which outperform existing state-of-the-art models and do not require heavy preprocessing.

A plethora of research has been done using feature-based approaches or methods to extract a set of features that represent time series patterns. Bag-of-Words (BoW) [15], Bag-of-features (TSBF) [16], Bag-of-SFA-Symbols (BOSS) [17], BOSSVS [18], and Word ExtrAction for time Series cLas-sification (WEASEL) [19] have obtained promising results in the field. Bag-of-words quantizes the extracted features and feeds the BoW into a classifier. TSBF extracts multiple subsequences of random local information, which a supervised learner condenses into a cookbook used to predict time series labels. BOSS introduces a combination of a distance

based classifier and histograms. The histograms represent substructures of a time series that are created using a symbolic Fourier approximation. BOSSVS extends this method by proposing a vector space model to reduce time complexity while maintaining performance. WEASEL converts time series into feature vectors using a sliding window. Machine learning algorithms utilize these feature vectors to detect and classify the time series. All these classifiers require heavy feature extraction and feature engineering. Using multiple of these feature-based algorithms as an ensemble algorithm yields better results.

Ensemble algorithms also yield state-of-the-art performance with time series classification problems. Three of the most successful ensemble algorithms that integrate various features of a time series are Proportional Elastic Ensemble (PROP) [20], a model that integrates 11 time series classifiers using a weighted ensemble method, shapelet ensemble (SE) [8], a model that applies a heterogeneous ensemble onto transformed shapelets, and a flat collective of transform based ensembles (COTE) [8], a model that fuses 35 various classifiers into a single classifier.

Recently, deep neural networks have been employed for time series classification tasks. Multi-scale convolutional neural network (MCNN) [12], fully convolutional

network (FCN) [11], and residual network (ResNet) [11] are deep learning approaches that take advantage of convolutional neural networks (CNN) for end-to-end classification of univariate time series. MCNN uses down-sampling, skip sampling and sliding window to preprocess the data. The performance of the MCNN classifier is highly dependent on the preprocessing applied to the dataset and the tuning of a large set of hyperparameters of that model. On the other hand, FCN and ResNet do not require any heavy preprocessing on the data or feature engineering.

In this paper, we improve the performance of FCN by augmenting the FCN module with either a Long Short Term Recurrent Neural Network (LSTM RNN) sub-module, called LSTM-FCN, or a LSTM RNN with attention, called ALSTM-FCN. In addition, the Attention LSTM can also be used to detect regions of the input sequence that contribute to the class label through the context vector of the Attention LSTM cells. Results indicate the new proposed models, LSTM-FCN and ALSTM-FCN, dramatically improve performance on the University of California Riverside (UCR) Benchmark datasets [21]. LSTM-FCN and ALSTM-FCN produce better results than several state-of-the-art algorithms on a majority of the UCR Benchmark datasets.

This paper proposes two deep learning models for end-to-end time series classification. The proposed models do not require heavy preprocessing on the data or feature engineering. Both the models are tested on all 85 UCR time series benchmarks and outperform most of the state-of-the-art models. The remainder of the paper is organized as follows. Section II reviews the background work. Section III presents the architecture of the proposed models. Section IV analyzes and discusses the experiments performed. Finally, conclusions are drawn in Section V.

II. BACKGROUND WORKS

A. TEMPORAL CONVOLUTIONS

The input to a Temporal Convolutional Network is generally a time series signal. As stated in Lea *et al.* [22], let $X_t \in \mathbb{R}^{F_0}$ be the input feature vector of length F_0 for time step t for $0 < t \leq T$. Note that the time T may vary for each sequence, and we denote the number of time steps in each layer as T_l . The true action label for each frame is given by $y_t \in \{1, \dots, C\}$, where C is the number of classes.

Consider L convolutional layers. We apply a set of 1D filters on each of these layers that capture how the input signals evolve over the course of an action. According to Lea *et al.* [22], the filters for each layer are parameterized by tensor $W^{(l)} \in \mathbb{R}^{F_l \times d \times F_{l-1}}$ and biases $b^{(l)} \in \mathbb{R}^{F_l}$, where $l \in \{1, \dots, L\}$ is the layer index and d is the filter duration. For the l -th layer, the i -th component of the (unnormalized) activation $\hat{\mathbf{E}}_{i,t}^{(l)} \in \mathbb{R}^{F_l}$ is a function of the incoming (normalized) activation matrix $E^{(l-1)} \in \mathbb{R}^{F_{l-1} \times T_{l-1}}$ from the previous layer

$$\hat{\mathbf{E}}_{i,t}^{(l)} = f \left(b_i^{(l)} + \sum_{t'=1}^d \langle W_{i,t',t}^{(l)}, E_{:,t+d-t'}^{(l-1)} \rangle \right) \quad (1)$$

for each time t where $f(\cdot)$ is a Rectified Linear Unit.

We use Temporal Convolutional Networks as a feature extraction module in a Fully Convolutional Network (FCN) branch. A basic convolution block consists of a convolution layer, followed by batch normalization [23], followed by an activation function, which can be either a Rectified Linear Unit or a Parametric Rectified Linear Unit [24].

B. RECURRENT NEURAL NETWORKS

Recurrent Neural Networks, often shortened to RNNs, are a class of neural networks which exhibit temporal behaviour due to directed connections between units of an individual layer. As reported by Pascanu *et al.* [25], recurrent neural networks maintain a hidden vector \mathbf{h} , which is updated at time step t as follows:

$$\mathbf{h}_t = \tanh(\mathbf{W}\mathbf{h}_{t-1} + \mathbf{I}\mathbf{x}_t), \quad (2)$$

\tanh is the hyperbolic tangent function, \mathbf{W} is the recurrent weight matrix and \mathbf{I} is a projection matrix. The hidden state \mathbf{h} is used to make a prediction

$$\mathbf{y}_t = \text{softmax}(\mathbf{W}\mathbf{h}_{t-1}), \quad (3)$$

softmax provides a normalized probability distribution over the possible classes, σ is the logistic sigmoid function and \mathbf{W} is a weight matrix. By using \mathbf{h} as the input to another RNN, we can stack RNNs, creating deeper architectures

$$\mathbf{h}_t^l = \sigma(\mathbf{W}\mathbf{h}_{t-1}^l + \mathbf{I}\mathbf{h}_t^{l-1}). \quad (4)$$

C. LONG SHORT-TERM MEMORY RNNs

Long short-term memory recurrent neural networks are an improvement over the general recurrent neural networks, which possess a vanishing gradient problem. As stated in Hochreiter and Schmidhuber [26], LSTM RNNs address the vanishing gradient problem commonly found in ordinary recurrent neural networks by incorporating gating functions into their state dynamics. At each time step, an LSTM maintains a hidden vector \mathbf{h} and a memory vector \mathbf{m} responsible for controlling state updates and outputs. More concretely, Graves *et al.* [27] define the computation at time step t as follows :

$$\begin{aligned} \mathbf{g}^u &= \sigma(\mathbf{W}^u \mathbf{h}_{t-1} + \mathbf{I}^u \mathbf{x}_t) \\ \mathbf{g}^f &= \sigma(\mathbf{W}^f \mathbf{h}_{t-1} + \mathbf{I}^f \mathbf{x}_t) \\ \mathbf{g}^o &= \sigma(\mathbf{W}^o \mathbf{h}_{t-1} + \mathbf{I}^o \mathbf{x}_t) \\ \mathbf{g}^c &= \tanh(\mathbf{W}^c \mathbf{h}_{t-1} + \mathbf{I}^c \mathbf{x}_t) \\ \mathbf{m}_t &= \mathbf{g}^f \odot \mathbf{m}_{t-1} + \mathbf{g}^u \odot \mathbf{g}^c \\ \mathbf{h}_t &= \tanh(\mathbf{g}^o \odot \mathbf{m}_t) \end{aligned} \quad (5)$$

where σ is the logistic sigmoid function, \odot represents elementwise multiplication, $\mathbf{W}^u, \mathbf{W}^f, \mathbf{W}^o, \mathbf{W}^c$ are recurrent weight matrices and $\mathbf{I}^u, \mathbf{I}^f, \mathbf{I}^o, \mathbf{I}^c$ are projection matrices.

While LSTMs possess the ability to learn temporal dependencies in sequences, they have difficulty with long term dependencies in long sequences. The attention mechanism proposed by Bahdanau *et al.* [28] can help the LSTM RNN learn these dependencies.

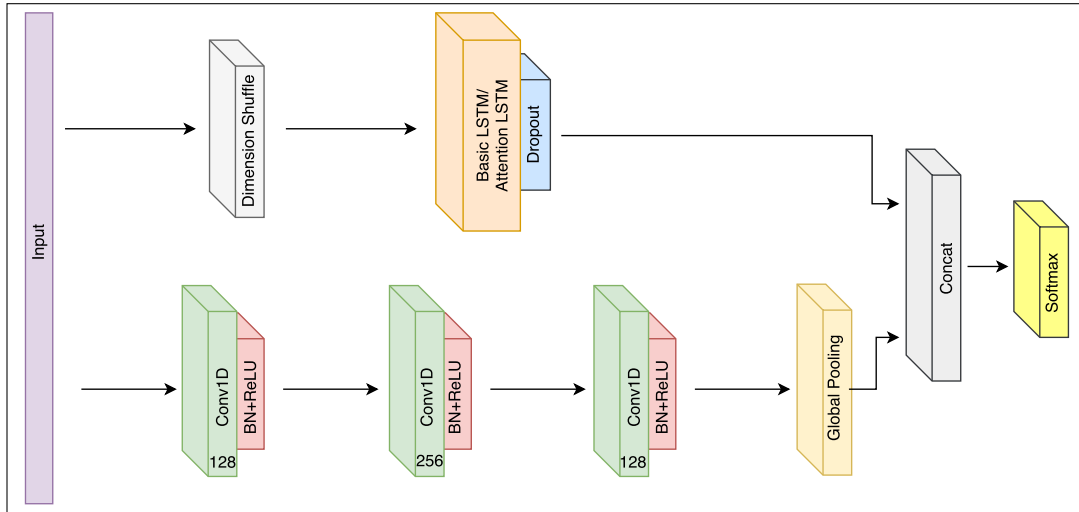


FIGURE 1. The LSTM-FCN architecture. LSTM cells can be replaced by Attention LSTM cells to construct the ALSTM-FCN architecture.

D. ATTENTION MECHANISM

The attention mechanism is a technique often used in neural translation of text, where a context vector C is conditioned on the target sequence y . As discussed in Bahdanau *et al.* [28], the context vector c_i depends on a sequence of annotations (h_1, \dots, h_{T_x}) to which an encoder maps the input sequence. Each annotation h_i contains information about the whole input sequence with a strong focus on the parts surrounding the i -th word of the input sequence.

The context vector c_i is then computed as a weighted sum of these annotations h_i :

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij} h_j. \quad (6)$$

The weight α_{ij} of each annotation h_j is computed by :

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})} \quad (7)$$

where $e_{ij} = a(s_{i-1}, h_j)$ is an *alignment model*, which scores how well the input around position j and the output at position i match. The score is based on the RNN hidden state $s_{i\hat{a}L_1}$ and the j -th annotation h_j of the input sentence.

Bahdanau *et al.* [28] parametrize the alignment model a as a feedforward neural network which is jointly trained with all the other components of the model. The alignment model directly computes a soft alignment, which allows the gradient of the cost function to be backpropagated.

III. LSTM FULLY CONVOLUTIONAL NETWORK

A. NETWORK ARCHITECTURE

Temporal convolutions have proven to be an effective learning model for time series classification problems [11]. Fully Convolutional Networks, comprised of temporal convolutions, are typically used as feature extractors. Global average

pooling [29] is used to reduce the number of parameters in the model prior to classification. In the proposed models, the fully convolutional block is augmented by an LSTM block followed by dropout [30], as shown in Figure 1.

The fully convolutional block consists of three stacked temporal convolutional blocks with filter sizes of 128, 256, and 128 respectively. Each convolutional block is identical to the convolution block in the CNN architecture proposed by Wang *et al.* [11]. Each block consists of a temporal convolutional layer, which is accompanied by batch normalization [23] (momentum of 0.99, epsilon of 0.001) and followed by a ReLU activation function. Finally, global average pooling is applied after the final convolution block.

Simultaneously, the time series input is conveyed into a dimension shuffle layer (explained more in Section III-B). The transformed time series from the dimension shuffle is then passed into the LSTM block. The LSTM block, comprising of either a general LSTM layer or an Attention LSTM layer, is followed by a dropout. The output of the global pooling layer and the LSTM block is concatenated and passed onto a softmax classification layer.

B. NETWORK INPUT

The fully convolutional block and LSTM block perceive the same time series input in two different views. The fully convolutional block views the time series as a univariate time series with multiple time steps. If there is a time series of length N , the fully convolutional block will receive the data in N time steps.

In contrast, the LSTM block in the proposed architecture receives the input time series as a multivariate time series with a single time step. This is accomplished by the dimension shuffle layer, which transposes the temporal dimension of the time series. A univariate time series of length N ,

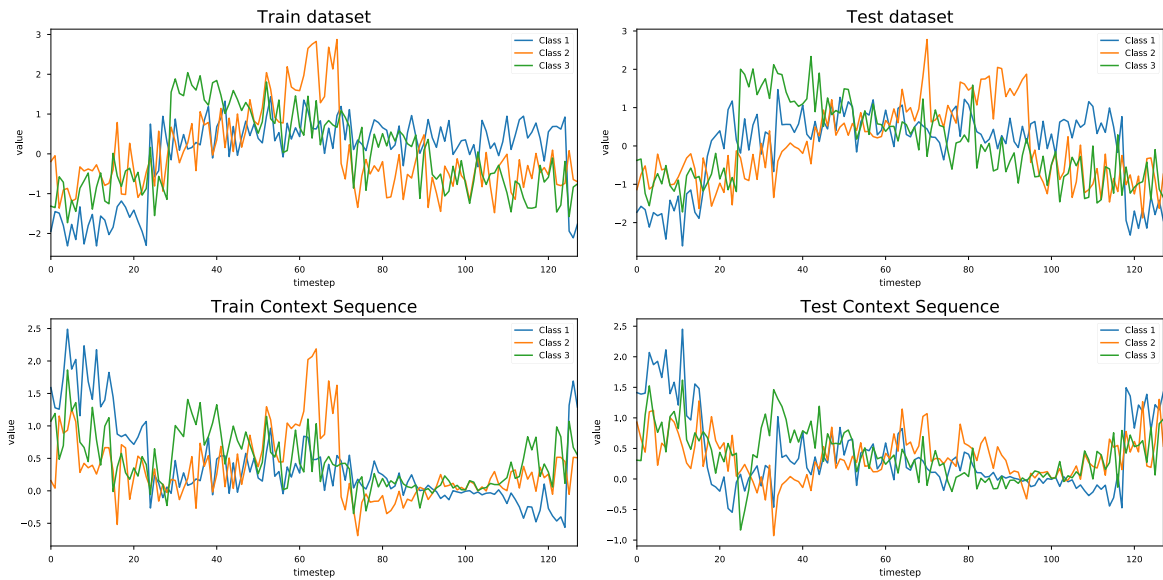


FIGURE 2. Visualization of context vector on CBF dataset.

Algorithm 1 Refinement

```

1: for  $i < K$  do
2:    $model_{weights} \leftarrow initial\_model_{weights}$ 
3:    $Train(model, initial\_lr, batchsize)$ 
4:    $initial\_model_{weights} \leftarrow model_{weights}$ 
5:    $i \leftarrow i + 1$ 
6:    $initial\_lr \leftarrow updateLearningRate(initial\_lr, i)$ 
7:    $batchsize \leftarrow updateBatchsize(batchsize, i)$ 

```

after transformation, will be viewed as a multivariate time series (having N variables) with a single time step. Without the dimension shuffle, the performance of the LSTM block is significantly reduced due to the rapid overfitting of small short-sequence UCR datasets and a failure to learn long term dependencies in the larger long-sequence UCR datasets.

In addition, dimension shuffle improves the efficiency of this model by requiring an order of magnitude less time to train. When a dataset of N time steps and M variables use a LSTM without dimension shuffling, the LSTM will require N time steps to process a batch of M variables. In contrast, applying the dimension shuffle to the input will allow the LSTM model to process a batch of N variables in M time steps. This suggests that as long as the number of variables M is significantly smaller than the number of time steps N , dimension shuffle will greatly improve the speed of training. As each of the UCR datasets is univariate, the LSTM component of this model will require only 1 time step to process a batch of N variables.

To illustrate this, a total of 18 hours is required on a single GTX 1080 Ti to train an LSTM-FCN for each of the 85 UCR datasets, and 19 hours for ALSTM-FCN.

Without the dimension shuffle, it would take more than 100 hours to train the respective models on all 85 UCR datasets.

C. REFINEMENT OF MODELS

Transfer learning is a technique wherein the knowledge gained from training a model on a dataset can be reused when training the model on another dataset, such that the domain of the new dataset has some similarity with the prior domain [31]. Similarly, we propose *refinement*, which can be described as transfer learning on the same dataset.

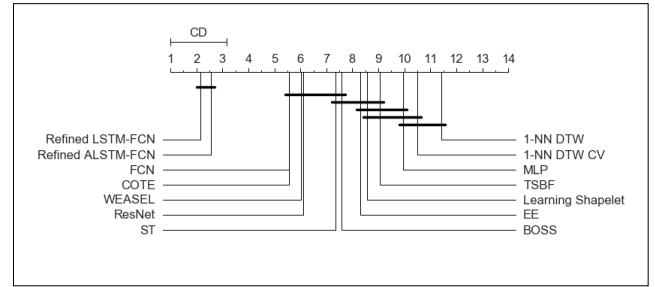
The training procedure can thus be split into two distinct phases. In the *initial phase*, the optimal hyperparameters for the model are selected for a given dataset. The model is then trained on the given dataset with these hyperparameter settings. In the second step, we apply *refinement* to this initial model.

The procedure of transfer learning is iterated over in the *refinement phase*, using the original dataset. Each repetition is initialized using the model weight of the previous iteration. At each iteration the learning rate is halved. Furthermore, the batch size is halved once every alternate iteration. This is done until the initial learning rate is $1e-4$ and batch size is 32. The procedure is repeated K times, where K is an arbitrary constant, generally set as 5.

Refinement is a procedure which successively attempts to improve the performance of a pre-trained model. As discussed by Huang et al. [32], multiple local minima lie along the optimization path of a model. Once a model has converged to some local minima during its initial training phase, it can be re-trained using a larger learning rate to escape the previous minima and hopefully land upon a better local minima.

TABLE 1. Performance comparison of proposed models with the rest.

Dataset	Existing SOTA	LSTM-FCN	Refined LSTM-FCN	ALSTM-FCN	Refined ALSTM-FCN
Adiac	0.8570 [11]	0.8593	0.8849	0.8670	0.8900*
ArrowHead	0.8800 [11]	0.9086	0.9029	0.9257*	0.9200
Beef	0.9000 [38]	0.9000	0.9330	0.9333*	0.9333*
BeetleFly	0.9500 [11, 19]	0.9500	1.0000*	1.0000*	1.0000*
BirdChicken	0.9500 [11, 17]	1.0000*	1.0000*	1.0000*	1.0000*
Car	0.9330 [11]	0.9500	0.9670	0.9667	0.9833*
CBF	1.0000 [11]	0.9978	1.0000*	0.9967	0.9967
ChioConc	0.8720 [11]	0.8099	1.0000*	0.8070	0.8070
CinC_ECG	0.9949 [8]	0.8862	0.9094	0.9058	0.9058
Coffee	1.0000 [19, 39]	1.0000*	1.0000*	1.0000*	1.0000*
Computers	0.8480 [11]	0.8600	0.8600	0.8640*	0.8640*
Cricket_X	0.8210 [11]	0.8077	0.8256*	0.8051	0.8051
Cricket_Y	0.8256 [8]	0.8179	0.8256*	0.8205	0.8205
Cricket_Z	0.8154 [8]	0.8103	0.8257	0.8308	0.8333*
DiaSizeRed	0.9670 [20]	0.9673	0.9771*	0.9739	0.9739
DistPhxAgeGp	0.8350 [11]	0.8600	0.8600	0.8625*	0.8600
DistPhxCorr	0.8200 [11]	0.8250	0.8217	0.8417*	0.8383
DistPhxTW	0.7900 [11]	0.8175	0.8100	0.8175	0.8200*
Earthquakes	0.8010 [11]	0.8354*	0.8261	0.8292	0.8292
ECG200	0.9200 [11]	0.9000	0.9200*	0.9100	0.9200
ECG5000	0.9482 [19]	0.9473	0.9478	0.9484	0.9496*
ECGFiveDays	1.0000 [19, 39]	0.9919	0.9942	0.9954	0.9954
ElectricDevices	0.7993 [17]	0.7681	0.7633	0.7672	0.7672
FaceAll	0.9290 [17]	0.9402	0.9680	0.9657	0.9728*
FaceFour	1.0000 [16, 19]	0.9432	0.9772	0.9432	0.9432
FacesUCR	0.9580 [11]	0.9293	0.9898*	0.9434	0.9434
FiftyWords	0.8198 [20]	0.8044	0.8066	0.8242	0.8286*
Fish	0.9890 [11]	0.9829	0.9886	0.9771	0.9771
FordA	0.9727 [19]	0.9272	0.9733*	0.9267	0.9267
FordB	0.9173 [39]	0.9180	0.9186*	0.9158	0.9158
Gun_Point	1.0000 [19, 38]	1.0000*	1.0000*	1.0000*	1.0000*
Ham	0.7810 [11]	0.7714	0.8000	0.8381*	0.8000
HandOutlines	0.9487 [19]	0.8930	0.8870	0.9030	0.9030
Haptics	0.5510 [11]	0.5747*	0.5584	0.5649	0.5584
Herring	0.7030 [11]	0.7656*	0.7188	0.7500	0.7656*
InlineSkate	0.6127 [19]	0.4655	0.5000	0.4927	0.4927
InsWngSnd	0.6525 [8]	0.6616	0.6696	0.6823*	0.6818
ITPdDmd	0.9700 [11]	0.9631	0.9699	0.9602	0.9708*
LrgKitApp	0.8960 [11]	0.9200*	0.9200*	0.9067	0.9120
Lighting2	0.8853 [20]	0.8033	0.8197	0.7869	0.7869
Lighting7	0.8630 [11]	0.8356	0.9178*	0.8219	0.9178*
Mallat	0.9800 [11]	0.9808	0.9834	0.9838	0.9842*
Meat	1.0000 [11]	0.9167	1.0000*	0.9833	1.0000*
MedicalImages	0.7920 [11]	0.8013	0.8066*	0.7961	0.7961
MidPhxAgeGp	0.8144 [16]	0.8125	0.8150	0.8175*	0.8075
MidPhxCorr	0.8076 [16]	0.8217	0.8333	0.8400	0.8433*
MidPhxTW	0.6120 [11]	0.6165	0.6466	0.6466*	0.6316
MoteStrain	0.9500 [11]	0.9393	0.9569*	0.9361	0.9361
NonInv_Thor1	0.9610 [11]	0.9654	0.9657	0.9751	0.9756*
NonInv_Thor2	0.9550 [11]	0.9623	0.9613	0.9664	0.9674*
OliveOil	0.9333 [19]	0.8667	0.9333	0.9333	0.9667*
OSULeaf	0.9880 [11]	0.9959*	0.9959*	0.9959*	0.9917
PhalCorr	0.8300 [11]	0.8368	0.8392*	0.8380	0.8357
Phoneme	0.3492 [8]	0.3776*	0.3602	0.3671	0.3623
Plane	1.0000 [19, 20]	1.0000*	1.0000*	1.0000*	1.0000*
ProxPhxAgeGp	0.8832 [38]	0.8927*	0.8878	0.8878	0.8927*
ProxPhxCorr	0.9180 [11]	0.9450*	0.9313	0.9313	0.9381
ProxPhxTW	0.8150 [17]	0.8350	0.8275	0.8375*	0.8375*
RefDev	0.5813 [38]	0.5813	0.5947*	0.5840	0.5840
ScreenType	0.7070 [11]	0.6693	0.7073	0.6907	0.6907
ShapeletSim	1.0000 [17, 19]	0.9722	1.0000*	0.9833	0.9833
ShapesAll	0.9183 [19]	0.9017	0.9150	0.9183	0.9217*
SmlKitApp	0.8030 [11]	0.8080	0.8133*	0.7947	0.8133*
SonyAIBOI	0.9850 [11]	0.9817	0.9967	0.9700	0.9983*
SonyAIBOII	0.9620 [11]	0.9780	0.9822*	0.9748	0.9790
StarlightCurves	0.9796 [8]	0.9756	0.9763	0.9767	0.9767
Strawberry	0.9760 [17]	0.9838	0.9864	0.9838	0.9865*
SwedishLeaf	0.9664 [19]	0.9792	0.9840	0.9856*	0.9856*
Symbols	0.9668 [17]	0.9839	0.9849	0.9869	0.9889*
Synth_Cntr	1.0000 [8, 20]	0.9933	1.0000*	0.9900	0.9900
TocSeg1	0.9737 [8]	0.9825	0.9912*	0.9868	0.9868
TocSeg2	0.9615 [17]	0.9308	0.9462	0.9308	0.9308
Trace	1.0000 [19, 20]	1.0000*	1.0000*	1.0000*	1.0000*
Two_Patterns	1.0000 [11, 20]	0.9968	0.9973	0.9968	0.9968
TwoLeadECG	1.0000 [18, 20]	0.9991	1.0000*	0.9991	1.0000*
uWayGest_X	0.8308 [16]	0.8490	0.8498	0.8481	0.8504*
uWayGest_Y	0.7585 [8]	0.7672*	0.7661	0.7658	0.7644
uWayGest_Z	0.7725 [16]	0.7973	0.7993	0.7982	0.8007*
uWayGestAll	0.9685 [20]	0.9618	0.9609	0.9626	0.9626
Wafer	1.0000 [19, 38]	0.9992	1.0000*	0.9981	0.9981
Wine	0.8890 [11]	0.8704	0.8890	0.9074*	0.9074*
WordsSynonyms	0.7790 [20]	0.6708	0.6991	0.6677	0.6677
Worms	0.8052 [19]	0.6685	0.6851	0.6575	0.6575
WormsTwoClass	0.8312 [38]	0.7956	0.8066	0.8011	0.8011
yoga	0.9183 [17]	0.9177	0.9163	0.9190	0.9237*
Count	-	43	65	51	57
MPCE	-	0.0318	0.0283	0.0301	0.0294
Arith. Mean	-	-	2.1529	-	2.5647
Geom. Mean	-	-	1.8046	-	1.8506

**FIGURE 3.** Critical difference diagram of the arithmetic means of the ranks.

IV. EXPERIMENTS

The proposed models have been tested on all 85 UCR time series datasets [21]. The FCN block was kept constant throughout all experiments. The optimal number of LSTM cells was found by hyperparameter search over a range of 8 cells to 128 cells. The number of training epochs was generally kept constant at 2000 epochs, but was increased for datasets where the algorithm required a longer time to converge. Initial batch size of 128 was used, and halved for each successive iteration of the *refinement* algorithm. A high dropout rate of 80% was used after the LSTM or Attention LSTM layer to combat overfitting. Class imbalance was handled via a class weighing scheme inspired by King and Zeng [33]. All models were trained using the Keras [34] library with the TensorFlow [35] backend.

All models were trained via the Adam optimizer [36], with an initial learning rate of $1e-3$ and a final learning rate of $1e-4$. All convolution kernels were initialized with the initialization proposed by He *et al.* [37]. The learning rate was reduced by a factor of $1/\sqrt{2}$ every 100 epochs of no improvement in the validation score, until the final learning rate was reached. No additional preprocessing was done on the UCR datasets as they have close to zero mean and unit variance. All models were *refined*. Scores stated in Table 1 refer to the scores obtained by models prior to and after *refinement*.¹

A. EVALUATION METRICS

In this paper, the proposed model was evaluated using accuracy, rank based statistics, and the mean per class error as stated by Wang *et al.* [11].

The rank-based evaluations used are the arithmetic rank, geometric rank, and the Wilcoxon signed rank test. The arithmetic rank is the arithmetic mean of the rank of dataset. The geometric rank is the geometric mean of the rank of each dataset. The Wilcoxon signed rank test is used to compare the median rank of the proposed model and the existing state-of-the-art models. The null hypothesis and alternative hypothesis are as follows:

$$H_0 : Median_{proposed\ model} = Median_{state-of-the-art\ model}$$

¹The codes and weights of each models are available at <https://github.com/houshd/LSTM-FCN> <https://github.com/houshd/LSTM-FCN>

Re-training while simultaneously reducing the learning rate and batch size allows for a more refined search to the best local optima.

TABLE 2. Wilcoxon signed rank test comparison of each model.

	WEASEL	1-NN DTW CV	1-NN DTW	BOSS	Learning Shapelet	TSBF	ST	EE	COTE	MLP	CNN	ResNet	LSTM- FCN	Refined LSTM- FCN	ALSTM- FCN
1-NN DTW CV	2.39E-10														
1-NN DTW	2.53E-12	7.20E-04													
BOSS	4.27E-03	1.82E-07	5.31E-11												
Learning Shapelet	2.00E-04	2.53E-02	2.33E-04	1.94E-02											
TSBF	2.18E-05	1.59E-01	2.49E-03	4.36E-03	4.73E-01										
ST	1.29E-01	1.05E-07	9.64E-11	2.39E-01	1.61E-03	3.60E-04									
EE	4.51E-05	3.45E-07	1.31E-10	1.37E-02	6.13E-01	2.02E-01	1.39E-03								
COTE	5.44E-01	3.05E-14	3.03E-16	6.21E-04	4.76E-07	1.13E-06	4.24E-03	3.54E-11							
MLP	2.56E-07	5.21E-01	3.41E-01	6.89E-05	1.44E-02	8.37E-02	6.76E-06	4.88E-03	2.84E-08						
FCN	2.77E-01	1.84E-10	2.14E-15	1.03E-03	3.65E-06	1.54E-06	8.85E-03	6.07E-06	4.82E-01	2.79E-09					
ResNet	5.67E-01	1.82E-10	5.95E-15	4.38E-03	1.32E-05	3.56E-06	2.47E-02	1.09E-05	9.61E-01	4.64E-08	2.52E-01				
LSTM-FCN	4.92E-06	1.92E-17	8.59E-21	3.00E-11	2.65E-12	4.04E-12	9.93E-13	5.14E-13	1.60E-07	1.61E-14	1.05E-07	4.91E-10			
Refined LSTM-FCN	1.23E-08	5.17E-19	5.77E-22	3.35E-13	2.20E-13	1.12E-13	3.44E-14	1.25E-14	2.81E-10	5.09E-16	3.35E-12	4.58E-15	7.53E-05		
ALSTM-FCN	1.34E-07	2.74E-18	5.14E-21	1.34E-12	3.38E-12	7.48E-13	7.11E-14	1.26E-13	1.30E-08	1.70E-15	3.74E-09	1.33E-11	8.53E-04	3.06E-02	
Refined ALSTM-FCN	4.58E-08	1.01E-18	1.18E-21	1.44E-12	2.41E-12	4.63E-13	3.96E-14	4.12E-14	2.56E-09	1.87E-15	2.60E-10	1.12E-12	5.96E-05	1.89E-01	5.40E-02

$$H_a : \text{Median}_{\text{proposed model}} \neq \text{Median}_{\text{state-of-the-art model}}$$

Mean Per Class Error (MPCE) is defined as the arithmetic mean of the per class error (PCE),

$$PCE_k = \frac{1 - \text{accuracy}}{\text{number of unique classes}}$$

$$MPCE = \frac{1}{K} \sum PCE_k.$$

B. RESULTS

Figure 2 is an example of the visual representation of the Attention LSTM cell on the “CBF” dataset. The points in the figure where the sequences are “squeezed” together are points at which all the classes have the same weight. These are the points in the time series at which the Attention LSTM can correctly identify the class. This is further supported by visual inspection of the actual time series. The squeeze points are points where each of the classes can be distinguished from each other, as shown in Figure 2.

The performance of the proposed models on the UCR datasets are summarized in Table 1. The colored cells are cells that outperform the state-of-the-art model for that dataset. Both proposed models, the ALSTM-FCN model and the LSTM-FCN model, with both phases, without *refinement* (Phase 1) and with *refinement* (Phase 2), outperforms the state-of-the-art models in at least 43 datasets. The average arithmetic rank in Figure 3 indicates the superiority of our proposed models over the existing state-of-the-art models. This is further validated using the Wilcoxon signed rank test, where the p-value of each of the proposed models are less than 0.05 when compared to existing state-of-the-art models, Table 2.

The Wilcoxon Signed Test also provides evidence that *refinement* maintains or improves the overall accuracy on each of the proposed models. The MPCE of the LSTM-FCN and ALSTM-FCN models was found to reduce by 0.0035 and 0.0007 respectively when *refinement* was applied. *Refinement* improves the accuracy of the LSTM-FCN models on a greater number of datasets as compared to the ALSTM-FCN models. We postulate that this discrepancy is due to the fact that the

²Green cells designate instances where our performance matches or exceeds state-of-the-art results. * denotes model with best performance.

TABLE 3. Summary of advantages of the proposed models.

Advantage	LSTM-FCN	Refined LSTM-FCN	ALSTM-FCN	Refined ALSTM-FCN
Performance		✓		✓
Visualization			✓	✓

LSTM-FCN model contains fewer total parameters than the ALSTM-FCN model. This indicates a lower rate of overfitting on the UCR datasets. As a consequence, *refinement* is more effective on the LSTM-FCN models for the UCR datasets.

A significant drawback of *refinement* is that it requires more training time due to the added computational complexity of re-training the model using smaller batch sizes. The disadvantages of *refinement* are mitigated when using the ALSTM-FCN within Phase 1. At the end of Phase 1, the ALSTM-FCN model outperforms the Phase 1 LSTM-FCN model. One of the major advantage of using the Attention LSTM cell is it provides a visual representation of the attention vector. The Attention LSTM also benefits from *refinement*, but the effect is less significant as compared to the general LSTM model. A summary of the performance of each model type on certain characteristics is provided on Table 3.

V. CONCLUSION & FUTURE WORK

With the proposed models, we achieve a notable improvement on the current state-of-the-art for time series classification using deep neural networks. Our baseline models, with and without refinement, are trainable end-to-end with nominal preprocessing and are able to achieve significantly improved performance. LSTM-FCNs are able to augment FCN models, appreciably increasing their performance with a nominal increase in the number of parameters. ALSTM-FCNs enable one to visually inspect the decision process of the LSTM RNN and provide a strong baseline on their own. *Refinement* can be applied as a general procedure to a model to further elevate its performance. The strong increase in performance in comparison to the FCN models shows that LSTM RNNs can beneficially supplement the performance of FCN modules for time series classification. An overall analysis of the performance of our model is provided and compared to other techniques.

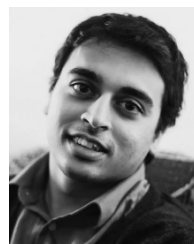
Due to the generality of the input to this model, it has wide ranging applicability on several sequence modelling tasks such as text analysis, music recognition and voice detection. Furthermore, due to its small size and efficiency, it can be easily deployed to real time systems or embedded systems. Additional research is to be done on understanding why the Attention LSTM cell is unsuccessful in matching the performance of the general LSTM cell on some of the datasets.

ACKNOWLEDGEMENT

Fazle Karim and Somshubra Majumdar contributed equally to this work.

REFERENCES

- [1] L. Wei and E. Keogh, "Semi-supervised time series classification," in *Proc. 12th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2006, pp. 748–753.
- [2] J. W. Taylor, P. E. McSharry, and R. Buizza, "Wind power density forecasting using ensemble predictions and time series models," *IEEE Trans. Sustain. Energy*, vol. 24, no. 3, pp. 775–782, Sep. 2009.
- [3] R. S. Tsay, *Analysis of Financial Time Series*, vol. 543. New York, NY, USA: Wiley, 2005.
- [4] L. C. Alwan and H. V. Roberts, "Time-series modeling for statistical process control," *J. Bus. Econ. Statist.*, vol. 6, no. 1, pp. 87–95, Jan. 1988.
- [5] A. T. Jebb, L. Tay, W. Wang, and Q. Huang, "Time series analysis for psychological research: Examining and forecasting change," *Frontiers Psychol.*, vol. 6, p. 727, Jun. 2015.
- [6] M. W. Kadous, "Temporal classification: Extending the classification paradigm to multivariate time series," Ph.D. dissertation, School Comput. Sci. Eng., Univ. New South Wales, Sydney, NSW, Australia, 2002.
- [7] A. Nanopoulos, R. Alcock, and Y. Manolopoulos, "Feature-based classification of time-series data," *Int. J. Comput. Res.*, vol. 10, no. 3, pp. 49–61, 2001.
- [8] A. Bagnall, J. Lines, J. Hills, and A. Bostrom, "Time-series classification with COTE: The collective of transformation-based ensembles," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 9, pp. 2522–2535, Sep. 2015.
- [9] A. Bagnall, J. Lines, J. Hills, and A. Bostrom, "Time-series classification with COTE: The collective of transformation-based ensembles," in *Proc. IEEE 32nd Int. Conf. Data Eng. (ICDE)*, May 2016, pp. 1548–1549.
- [10] B. Koley and D. Dey, "An ensemble system for automatic sleep stage classification using single channel EEG signal," *Comput. Biol. Med.*, vol. 42, no. 12, pp. 1186–1195, 2012.
- [11] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 1578–1585.
- [12] Z. Cui, W. Chen, and Y. Chen. (2016). "Multi-scale convolutional neural networks for time series classification." [Online]. Available: <https://arxiv.org/abs/1603.06995>
- [13] H. Wang, W. Sun, and P. X. Liu, "Adaptive intelligent control of nonaffine nonlinear time-delay systems with dynamic uncertainties," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 47, no. 7, pp. 1474–1485, Jul. 2017.
- [14] X. Zhao, P. Shi, X. Zheng, and J. Zhang, "Intelligent tracking control for a class of uncertain high-order nonlinear systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 9, pp. 1976–1982, Sep. 2016.
- [15] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing SAX: A novel symbolic representation of time series," *Data Mining Knowl. Discovery*, vol. 15, no. 2, pp. 107–144, Oct. 2007.
- [16] M. G. Baydogan, G. Runger, and E. Tuv, "A bag-of-features framework to classify time series," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2796–2802, Nov. 2013.
- [17] P. Schäfer, "The BOSS is concerned with time series classification in the presence of noise," *Data Mining Knowl. Discovery*, vol. 29, no. 6, pp. 1505–1530, Nov. 2015.
- [18] P. Schäfer, "Scalable time series classification," *Data Mining Knowl. Discovery*, vol. 30, no. 5, pp. 1273–1298, Sep. 2016.
- [19] P. Schäfer and U. Leser. (2017). "Fast and accurate time series classification with WEASEL." [Online]. Available: <https://arxiv.org/abs/1701.07681>
- [20] J. Lines and A. Bagnall, "Time series classification with ensembles of elastic distance measures," *Data Mining Knowl. Discovery*, vol. 29, no. 3, pp. 565–592, May 2015.
- [21] Y. Chen et al. (Jul. 2015). *The UCR Time Series Classification Archive*. [Online]. Available: http://www.cs.ucr.edu/~eamonn/time_series_data/
- [22] C. Lea, R. Vidal, A. Reiter, G. D. Hager, "Temporal convolutional networks: A unified approach to action segmentation," in *Computer Vision—ECCV (Lecture Notes in Computer Science)*, vol. 9915, G. Hua and H. Jégou, Eds. Cham, Switzerland: Springer, 2016.
- [23] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
- [24] L. Trottier, P. Giguère, and B. Chaib-draa. (May 2016). "Parametric exponential linear unit for deep convolutional neural networks." [Online]. Available: <http://arxiv.org/abs/1605.09332>
- [25] R. Pascanu, C. Gulcehre, K. Cho, and Y. Bengio. (2013). "How to construct deep recurrent neural networks." [Online]. Available: <https://arxiv.org/abs/1312.6026>
- [26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [27] A. Graves, *Supervised Sequence Labelling With Recurrent Neural Networks*, vol. 385. London, U.K.: Springer, 2012.
- [28] D. Bahdanau, K. Cho, and Y. Bengio. (2014). "Neural machine translation by jointly learning to align and translate." [Online]. Available: <https://arxiv.org/abs/1409.0473>
- [29] M. Lin, Q. Chen, and S. Yan. (2013). "Network in network." [Online]. Available: <https://arxiv.org/abs/1312.4400>
- [30] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jun. 2014.
- [31] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3320–3328.
- [32] G. Huang, Y. Li, G. Pleiss, Z. Liu, J. E. Hopcroft, and K. Q. Weinberger, "Snapshot Ensembles: Train 1, get M for free," *CoRR*, 2017. [Online]. Available: <http://arxiv.org/abs/1704.00109>
- [33] G. King and L. Zeng, "Logistic regression in rare events data," *Political Anal.*, vol. 9, no. 2, pp. 137–163, 2001.
- [34] F. Chollet et al. (2015). *Keras*. [Online]. Available: <https://github.com/fchollet/keras>
- [35] M. Abadi et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. [Online]. Available: <https://www.tensorflow.org/>
- [36] D. P. Kingma and J. Ba. (2014). "Adam: A method for stochastic optimization." [Online]. Available: <https://arxiv.org/abs/1412.6980>
- [37] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1026–1034.
- [38] A. Bostrom and A. Bagnall, "Binary shapelet transform for multiclass time series classification," in *Transactions on Large-Scale Data- and Knowledge-Centered Systems XXXII (Lecture Notes in Computer Science)*, vol. 10420, A. Hameurlain, J. Küng, R. Wagner, S. Madria, and T. Hara, Eds. Berlin, Germany: Springer, 2017.
- [39] J. Grabocka, N. Schilling, M. Wistuba, and L. Schmidt-Thieme, "Learning time-series shapelets," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 392–401.



FAZLE KARIM received the B.Sc. degree in industrial engineering from the University of Illinois at Urbana–Champaign in 2012, the M.Sc. degree in industrial engineering from the University of Illinois at Chicago in 2016. He is currently pursuing the Ph.D. degree with the Mechanical and Industrial Engineering Department, University of Illinois at Chicago. He is also the Lead Data Scientist with Prominent Laboratory, the university's foremost research facility in process mining. His research interests include education data mining, health care data mining, and time series analysis.



SOMSHUBRA MAJUMDAR received the B.S. degree in computer engineering from the University of Mumbai in 2016. He is currently pursuing the M.S. degree in computer science with the University of Illinois at Chicago. He is currently an aspiring Artificial Intelligence Researcher. His research interests lie in the domain of image classification and segmentation using convolutional neural networks, time series classification using recurrent neural networks, and machine learning.



SHUN CHEN received the B.S. and M.S. degrees in mathematics from Southeast University, Nanjing, China, in 2009 and 2012, respectively, and the Ph.D. degree in mathematics from the City University of Hong Kong in 2015. He was an Associate Professor with the Huazhong Agricultural University and Chutian Scholar of Hubei Province, China. He is currently a Post-Doctoral Researcher with the University of Illinois at Chicago. His research interests include time series, data mining,

fault analysis, multi-agent systems, complex dynamical networks, and sensor networks.

...



HOUSHANG DARABI (S'98–A'00–M'10–SM'14) received the Ph.D. degree in industrial and systems engineering from Rutgers University, New Brunswick, NJ, in 2000. He is currently an Associate Professor with the Department of Mechanical and Industrial Engineering, University of Illinois at Chicago, where he also holds an Associate Professor position with the Department of Computer Science. His main research interests include the application of data mining, process

mining, and optimization in design and analysis of manufacturing, business, project management, and workflow management systems.

His research has been supported by several federal and private agencies, such as the National Science Foundation, National Institute of Standard and Technology, Department of Energy, and Motorola. He has published in different prestigious journals and conference proceedings, such as the IEEE TRANSACTIONS ON ROBOTICS and AUTOMATION, the IEEE TRANSACTIONS ON AUTOMATION SCIENCE and ENGINEERING, the IEEE TRANSACTIONS ON SYSTEMS, MAN and CYBERNETICS, and *Information Sciences*. He has been a contributing author of two books in the area of scalable enterprise systems and reconfigurable discrete event systems. He has extensively published on various automation and project management subjects including wireless sensory networks for location sensing, planning, and management of projects with tasks requiring multi-mode resources, and workflow modeling and management.