PREDICTION OF LUNG CANCER PATIENT SURVIVAL TIME VIA SUPERVISED MACHINE LEARNING TECHNIQUE

TGDK Sumanathilaka B150413CS
Eva Reshma Toppo B150129CS
Rakhee Poonam Lakra B150547CS
Aravind A B150263CS

Department of Computer Science and Engineering
NIT Calicut

2 APRIL, 2019

Outline

Abstract

- Early detection of cancer is essential for a rapid response and better chances of cure.
- Unfortunately, as the symptoms of the disease at the beginning are absent, early detection of symptoms is very difficult.
- Thus, it is necessary to discover and interpret new knowledge to prevent and minimize the risk adverse consequences.
- Machine learning is widely used in bioinformatics and particularly in lung cancer diagnosis.

Abstract

- The prediction is treated like a continuous target, rather than a classification into categories.
- To understand this problem more precisely, tools are needed to help oncologists to choose the treatment required for healing or prevention of recurrence by reducing the harmful effects of certain treatments and their costs.

Problem Statement

- Predicting the survival time of the patient suffering from lung cancer using various machine learning technique
  1. Multilayer Perceptron (MLP)
  2. Deep Belief Network (DBN)
  3. Long Short-Term Memory (LSTM)

Literature Survey

- Base Paper : Prediction of lung cancer patient survival via supervised machine learning classification techniques.
- Chip M. Lyncha, Behnaz Abdollahib, Joshua D. Fuquac, Alexandra R. de Carloc,James A. Bartholomaic, Rayeanne N. Balgemannc, Victor H. van Berkeld, Hermann B. Frieboesc
- International journal of Medical Informatics [2017]

Literature Survey

Models implemented

1. Linear Regression
2. Gradient Boost Machine
3. Decision Tree
4. Random Forest
5. Support Vector Machine
6. Ensemble

Literature Survey - Results

Model	RMSE	Standard Deviation	SD of residuals	Mean
Linear Regression	15.492	10.171	15.490	14.992
GBM	16.328	4.323	16.331	15.171
Decision Tree	15.252	10.786	15.264	15.253
Random Forest	15.197	10.453	15.199	15.075
SVM	18.073	4.988	16.867	8.697
Ensemble	15.811	5.967	5.967	14.893

Figure: Comparison of modeling techniques based on RMSE values. Both the standard deviation of the predictions and the standard deviation of the difference between predictions and the actual values (Standard Deviation of Residuals) are shown along with the ensemble weighting factors.

Brief Idea about the Dataset

Dataset

- Dataset is taken from SEER(Surveillance, Epidemiology, and End Results).

Selected SEER attributes and their respective descriptors. AJCC: American Joint Committee on Cancer

Number	Attribute	Description	Type
1	Age	Age at time of diagnosis.	Discrete
2	Grade	Appearance of cancer cells and how fast they may grow.	Numeric
3	Radiation Sequence with Surgery	Order of surgery and radiation therapy administered for patients who received both.	Numeric
4	Number of Primaries	Number of malignant tumors other than lung.	Discrete
5	T	AJCC component describing tumor size.	Numeric
6	N	AJCC component describing lymph node involvement.	Numeric
7	M	AJCC component describing tumor dissemination to other organs.	Numeric
8	Radiation	Indication of whether patient has received radiation.	Numeric
9	Stage	Stage of tumor – based on T, N, and M.	Numeric
10	Primary Site	Location of tumor within the lungs.	Numeric
11	First Malignant Primary Indicator	Based on cancers reported in SEER database for patient.	Numeric
12	Sequence Number	Order of lung cancer occurrence with respect to other cancers for this patient.	Discrete
13	CS Lymphnodes	Number of lymph nodes involved.	Numeric
14	Histology Recode – Broad Groupings	Microscopic composition of cells and/or tissues for specific primary. Used for staging and treatment determination.	Numeric
15	RXSumm – ScopeRegLNSur(2003+)	(Scope of Regional Lymph Node Surgery) – Procedure of removal, biopsy, or aspiration of regional lymph nodes.	Numeric
16	RXSumm – SurgPrimSite(1998+)	(Surgery of Primary Site) – Procedure to remove or destroy tissue of the primary site.	Numeric
17	DerivedSS1977	This item is derived "SEER Summary Stage 1977" from the Collaborative Stage (CS) algorithm, effective with 2004+ diagnosis.	Numeric
18	TumorSize	Measurement of tumor size.	Numeric
19	Survival Time	Number of months that patient is alive from date of diagnosis.	Discrete

Dataset

Screenshot of the database

1	7	72	2	0	2	40	20	10	0	70	3	0	2	2	2	0	0	7	72
2	30	66	3	2	3	40	20	0	1	53	1	1	2	2	2	5	46	4	72
3	55	92	3	0	4	40	20	0	0	53	3	0	3	2	2	0	0	7	72
4	19	76	3	0	2	30	10	0	0	52	3	0	2	1	2	5	56	7	72
5	12	76	3	0	2	20	10	10	1	70	1	0	2	1	1	0	0	7	72
6	5	73	4	0	2	30	20	0	1	52	1	0	2	2	2	0	0	3	72
7	3	54	3	0	2	40	30	10	1	70	3	0	2	6	8	0	0	7	72
8	1	57	3	0	1	40	20	10	0	70	3	1	0	2	0	0	0	7	72
9	0	80	3	0	1	20	20	10	1	70	1	1	0	2	2	0	0	7	72
10	8	80	3	3	1	40	20	0	1	53	1	1	0	2	1	4	0	4	72
11	15	85	3	0	1	20	20	0	0	52	1	1	0	2	2	0	0	4	72
12	5	79	3	0	1	40	20	10	1	70	1	1	0	2	5	0	0	7	72
13	2	73	4	0	1	40	23	11	0	88	3	1	0	1	16	3	33	4	72
14	8	53	3	0	1	20	20	0	1	52	1	1	0	2	1	0	0	3	72
15	6	75	1	0	1	20	20	0	1	52	1	1	0	2	5	0	0	4	72
16	0	59	3	0	1	40	20	10	0	70	1	1	0	2	1	0	0	7	72
17	11	83	3	0	1	20	20	10	0	70	3	1	0	2	1	0	0	7	72

Preprocessing

- According to the base paper,19 attributes were seperated from the dataset with 134 attributes.
- Preprocessing includes:
  1. Removal of missing values.
  2. Survival Time limited to a range of 0-72 months
  3. Removal of Noise

Models

1. Multilayer Perceptron (MLP)
2. Deep Belief Networks (DBNs)
3. Long Short-Term Memory (LSTM)

Models

Experimental Framework

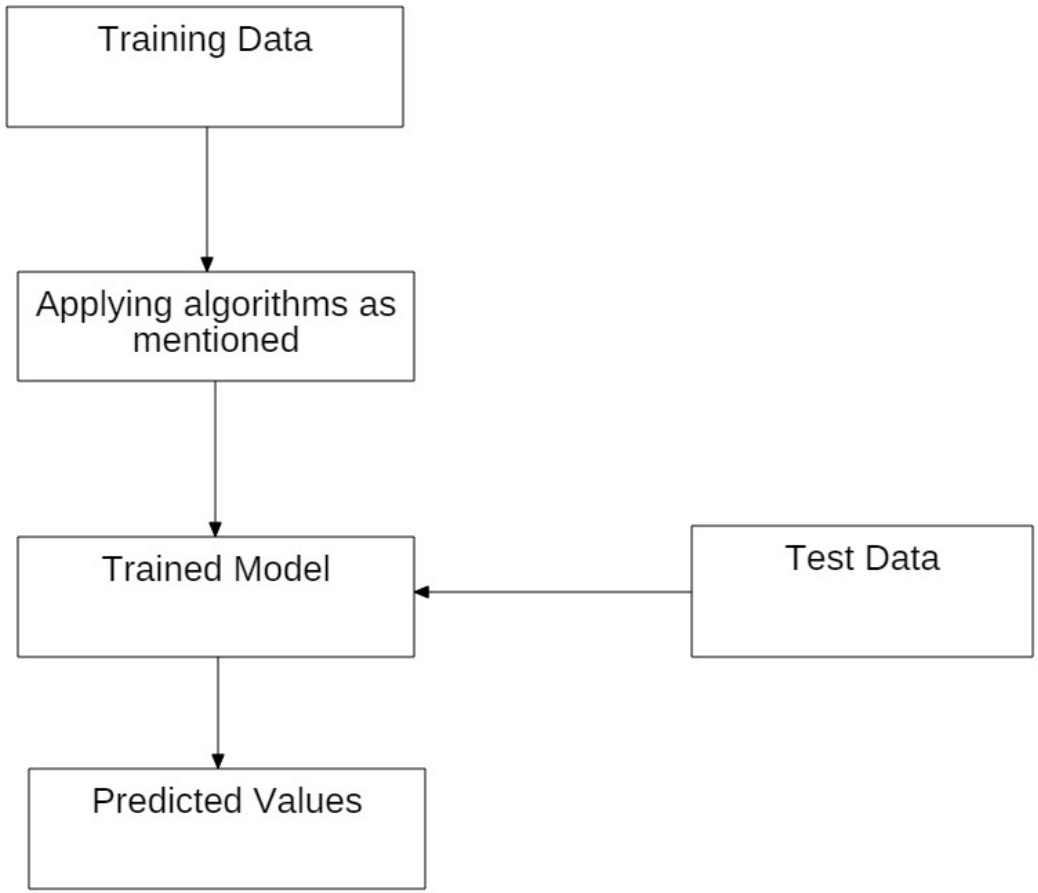


Figure: Flow Diagram of the Framework

Outline

Models

Multi Layer Perceptron(MLP)

- MLP is a class of feedforward artificial neural network.
- Consists of three layers of nodes:
  - An input layer
  - A hidden layer
  - An output layer
- MLP utilizes a supervised learning technique called backpropagation for training.
- Its multiple layers and non-linear activation distinguish MLP from a linear perceptron.

Multi Layer Perceptron (MLP)

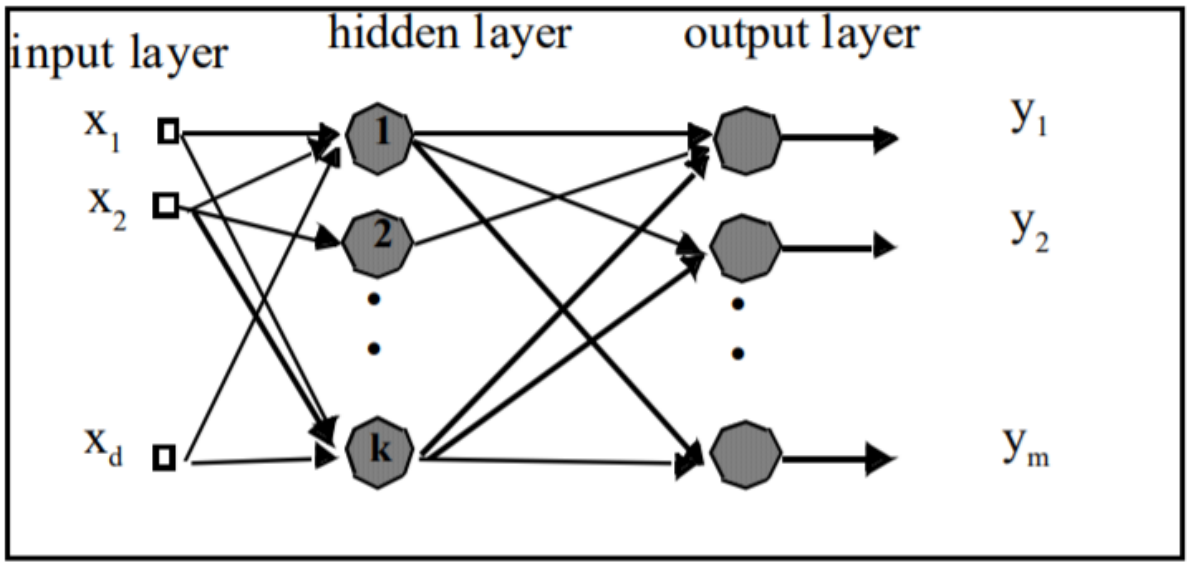


Figure: A multilayer perceptron with one hidden layer (d-k-m)

Multi Layer Perceptron (MLP)

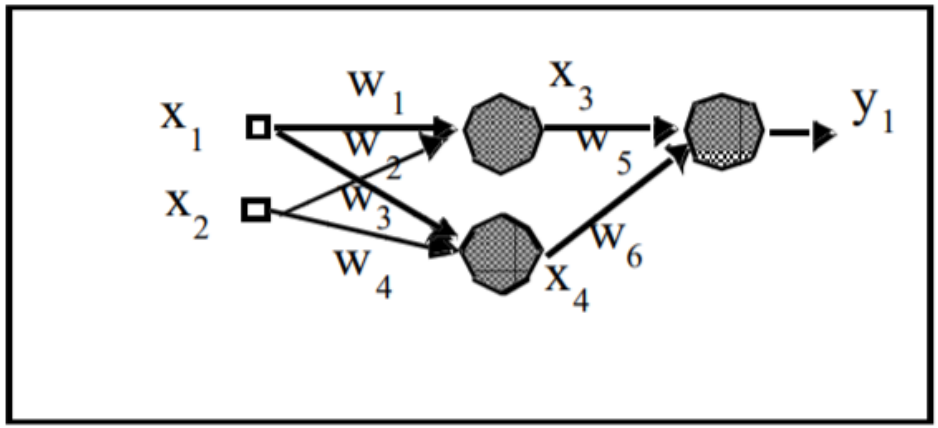


Figure: A multilayer perceptron with one hidden layer (2-2-1)

$$y = f\left(w_5 x_3 + w_6 x_4 + b_3\right) = f\left\{w_5\left[f\left(w_1 x_1 + w_2 x_2 + b_1\right)\right]\right.$$
$$\left. + w_6\left[f\left(w_3 x_1 + w_4 x_2 + b_2\right)\right] + b_3\right\} = f\left\{g_1 + g_2 + b_3\right\}$$

Models

Implementation

- Data set was split into .75 as training set and 0.25 as test data set
- Following libraries were used.
  - sklearn.neural.network - MLP Classifier
  - sklearn.preprocessing - for preprocessing the dataset
  - numpy - inbuilt arithmetic functions
  - pandas - for reading the csv file
  - math - for arithmetic calculations
- Maximum iteration 500

Models

Multi Layer Perceptron(MLP)

```
RMSE : 14.968406909479805
Mean of predictions :  46.32054077627562
Mean of residuals :  9.445268207588311
Standard deviation :  11.614600495054002
```

Figure: Output of MLP with 1 layer

```
RMSE : 14.878736470323556
Mean of predictions :  45.463148713475796
Mean of residuals :  9.382032272132577
Standard deviation :  11.550430809671578
```

Figure: Output of MLP with 2 layer

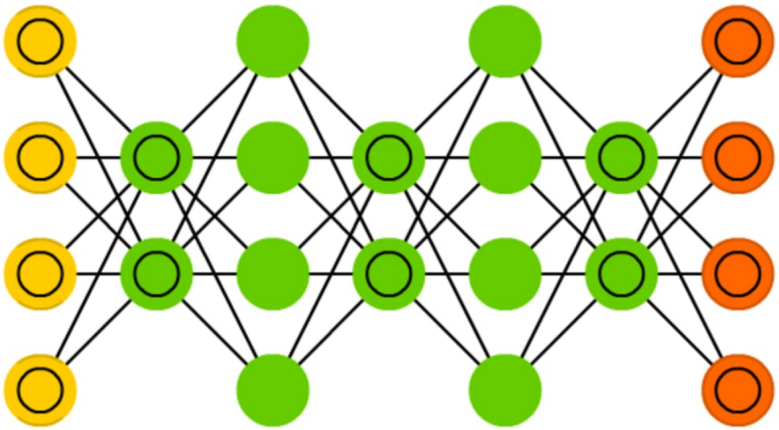
Outline

Models

Deep Belief Networks (DBN)

- A generative graphical model, or alternatively a class of deep neural network.
- Composed of multiple layers of latent variables ("hidden units"), with connections between the layers, but not between units within each layer
- Can be viewed as a composition of simple, unsupervised networks such as restricted Boltzmann machines (RBMs) or auto-encoders, where each sub-network's hidden layer serves as the visible layer for the next layer.

Deep Belief Network (DBN)



Models

Implementation

- Data set was split into .75 as training set and 0.25 as test data set
- Following libraries were used.
  - dbn.tensorflow
  - sklearn.preprocessing - for preprocessing the dataset
  - numpy - inbuilt arithmetic functions
  - pandas - for reading the csv file
  - math - for arithmetic calculations
- 50 epochs were used

Models

Deep Belief Networks(DBN)

```
RMSE : 16.399677169128527
Mean of predictions :  40.0
Mean of residuals :  14.590056694286961
Standard deviation :  7.490268651605505
```

Figure: Output of DBN

Outline

Models

Long Short-Term Memory(LSTM)

- LSTM is an artificial recurrent neural network (RNN) architecture used in the field of deep learning.
- A LSTM unit is composed of a cell, an input gate, an output gate and a forget gate.
- The cell remembers values over arbitrary time intervals and the three gates regulate the flow of information into and out of the cell.
- LSTM networks are well-suited to classifying, processing and making predictions based on time series data.

Models

Long Short-Term Memory(LSTM)

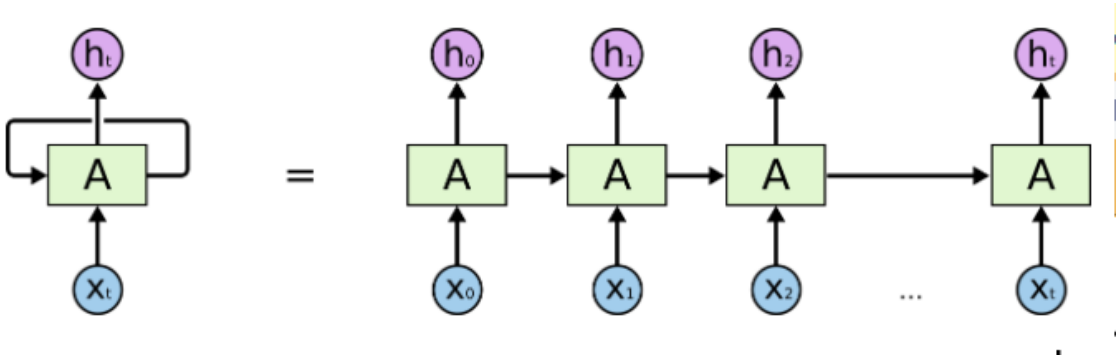


Figure: An unrolled recurrent neural network

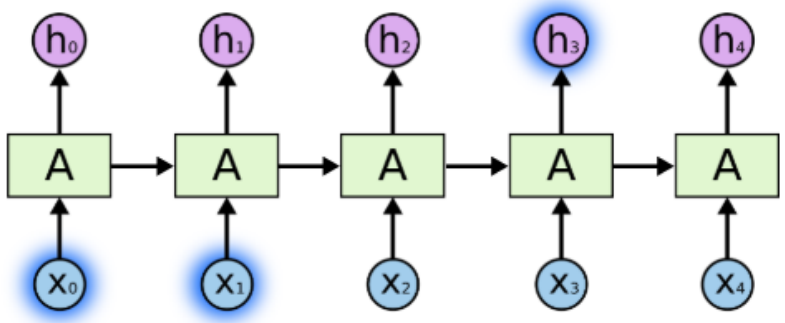


Figure: Recurrent Neural Netwrok

Models

Long Short-Term Memory(LSTM)

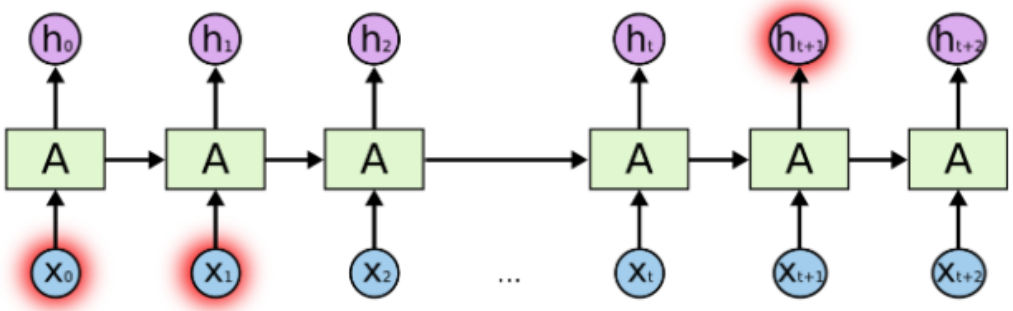


Figure: LSTM

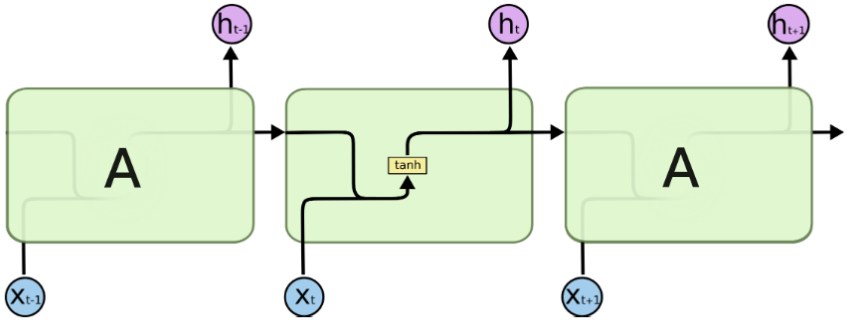


Figure: The repeating module in a standard RNN contains a single layer.

Models

Long Short-Term Memory(LSTM)

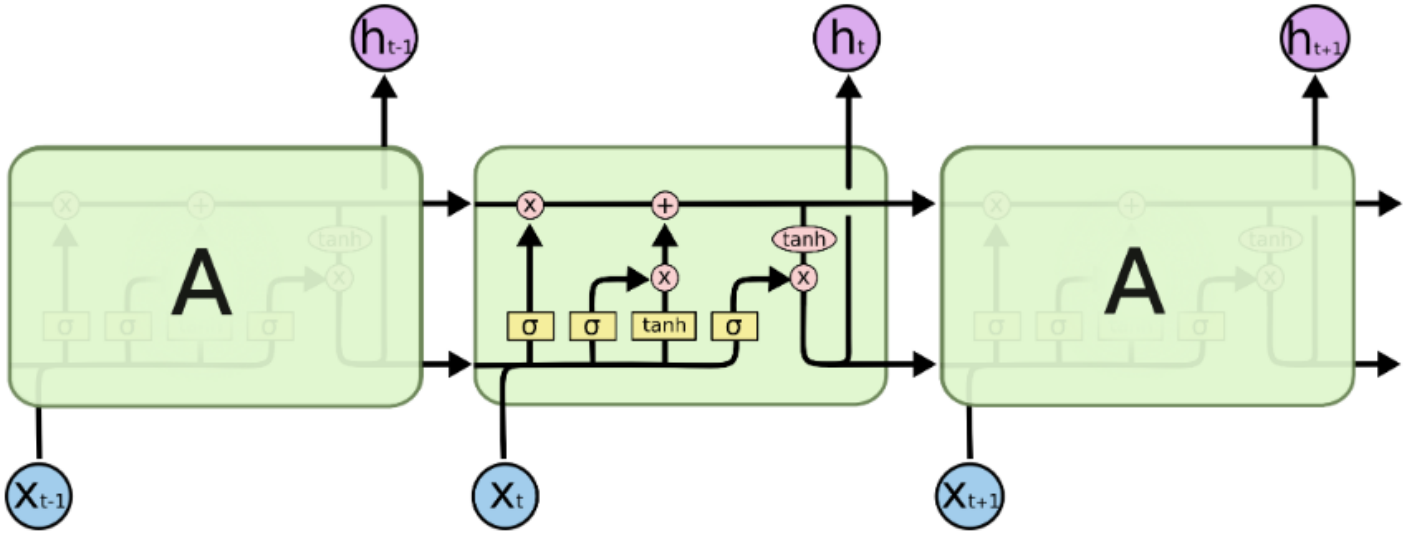


Figure: The repeating module in an LSTM contains four interacting layers.

Models

Long Short-Term Memory(LSTM)

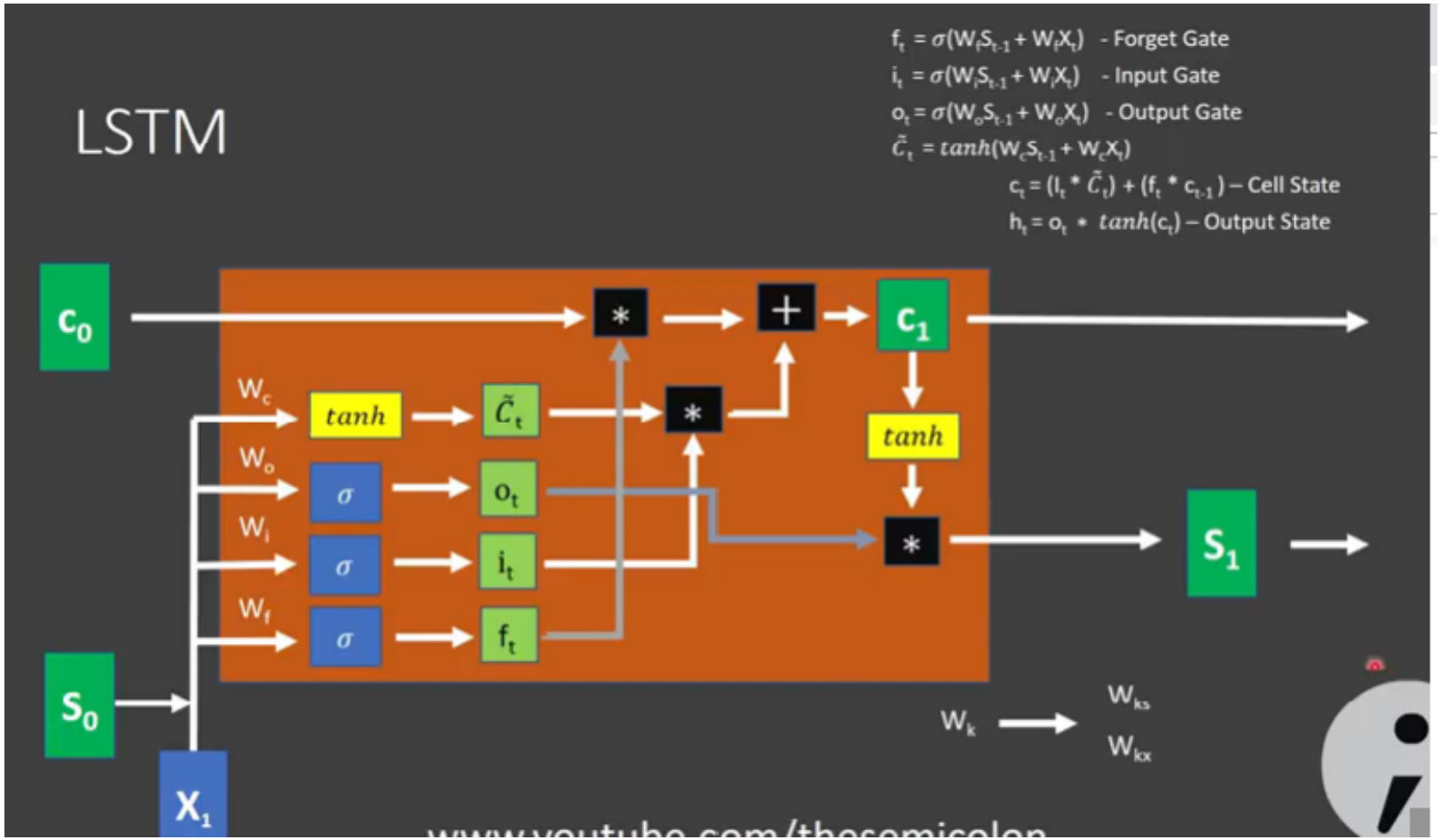


Figure: LSTM

Models

Implementation

- Data set was split into .75 as training set and 0.25 as test data set
- Following libraries were used.
  - Keras (LSTM)
  - sklearn.preprocessing - for preprocessing the dataset
  - numpy - inbuilt arithmetic functions
  - pandas - for reading the csv file
  - math - for arithmetic calculations
- 100 epochs were used

Models

Long Short-Term Memory(LSTM)

```
Mean Square root Error:10.53
Mean of predictions : 42.851722634103
Mean of residuals : 7.526412166
Standard deviation :14.26523564
```

Figure: Output of LSTM

Conclusion

MODEL	RMSE
LSTM	10.53
MLP with 1 layer	14.9684
MLP with 2 layer	14.8787
DBN	16.399

Table: Training Models and their RMSE values

- **LSTM** has the least **RMSE** value.
- Meaningful predictions can be made with **reasonable accuracy**.

REFERENCE I


Chip M. Lynch, Behnaz Abdollahi et.al
Prediction of lung cancer patient survival via supervised machine learning classification techniques, International Journal of Medical Informatics 108 (2017) 18.


Wikipedia
https://en.wikipedia.org/wiki/Long/*short* − *term*/*memory*


Wikipedia
http://colah.github.io/posts/2015-08-Understanding-LSTMs/


Wikipedia
https://en.wikipedia.org/wiki/Deep/*belief*/*network*

Thank You

CODE

You can find the code in the following link:

https://github.com/Sumanathilaka/Survival-Time-Prediction-of-a-patient-using-LSTM-MLP-DBN