ECE777AB FA23

STORAGE MEDIA TREND FAILURE ANALYSIS

BASED ON ENTERPRISE WORKLOADS AND ENVIRONMENTS

Md Reza E Rabbi R844H794

Yoel Woldeyes J479Q742

Sean Cowley V688J944

WICHITA STATE UNIVERSITY

INTRODUCTION

- Storage media holding critical information for businesses across all industries.
- Enterprise workloads demand high performance, reliability, and availability from storage systems.
- Storage media failures can lead to data loss, drive replacements, and productivity loss.

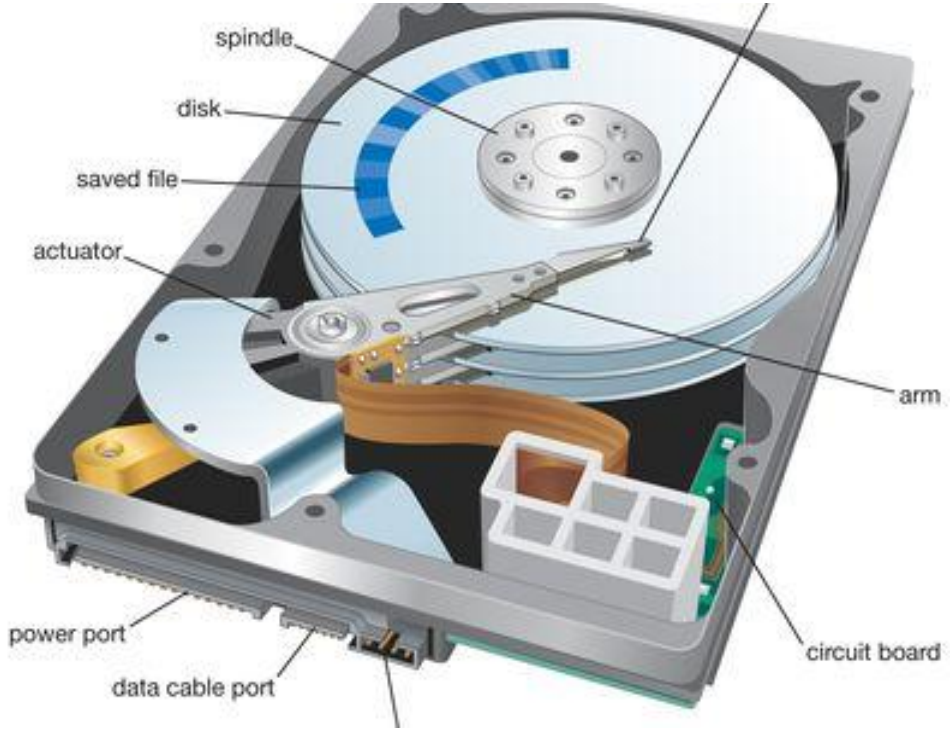


Fig 1: Hard Drive [1]

1. https://www.britannica.com/technology/hard-disk

PROBLEM STATEMENT

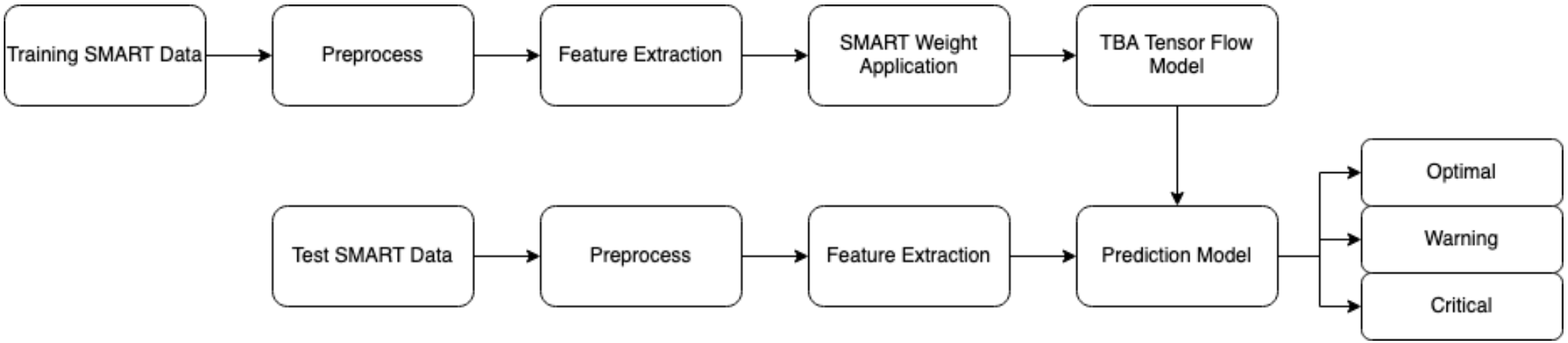
- Conventional methods such as Annualized Failure Rate (AFR), rely on post-failure analysis.
- Reactive approaches often lead to unexpected
  - data loss
  - downtime
  - increased costs.

RELATED WORK

- Transmitted vibration from speaker and fan can introduce PES. [2]
- Used FEM model and proposed chassis design and damping ratio to suppress vibration [2-3]
- Data-center environment influences HDD's performance by temperature and proposed a design for data-center to reduce temperature [4]
- By using "XGBoost", age of the HDD can monitor HDD's performance [5]

2. Y. Y. Hu, S. Yoshida- S. Nakamura, K. Watanabe, W. Z. Lin, E. T. Ong, and J. Q. Mou, "Analysis of built-in speaker-induced structural-acoustic vibration of hard disk drives in Notebook PCs" IEEE Trans. Magn. 2009.
3. J. Q. Mou, F. Lai, I. B. L. See and W. Z. Lin, "Analysis of structurally transmitted vibration of HDD in notebook computer," 2012 Digest APMRC, Singapore,
4. J. Zhao et al., "Disk Failure Early Warning Based on the Characteristics of Customized SMART", 2020 19th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2020.
5. Z. Miller, O. Medaiyese, M. Ravi, A. Beatty and F. Lin, "Hard Disk Drive Failure Analysis and Prediction: An Industry View," 2023 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks - Supplemental Volume (DSN-S), Porto Portugal, 2023

FLOWCHART


Training SMART Data
Preprocess
Feature Extraction
SMART Weight Application
TBA Tensor Flow Model
Test SMART Data
Preprocess
Feature Extraction
Prediction Model
Optimal
Warning
Critical

PROPOSED WORK

- **Objective**: Analyze failure trends in storage media within enterprise data centers.
- **Motivation**: Address the exponential demand for data storage driven by Artificial Intelligence growth and the need for effective storage solutions in large data centers.
- **Methodology**: Process and analyze data center failure reports from BackBlaze to identify trends in storage media failures.
- **Focus Areas**:
  - Variety of storage media vendors, models, and technologies.
  - Detection of trending features in failing storage media and models.
  - Root cause analysis of observed failures.
- **Validation and Testing**: Compare and validate findings using untrained yearly quarter results from data centers.

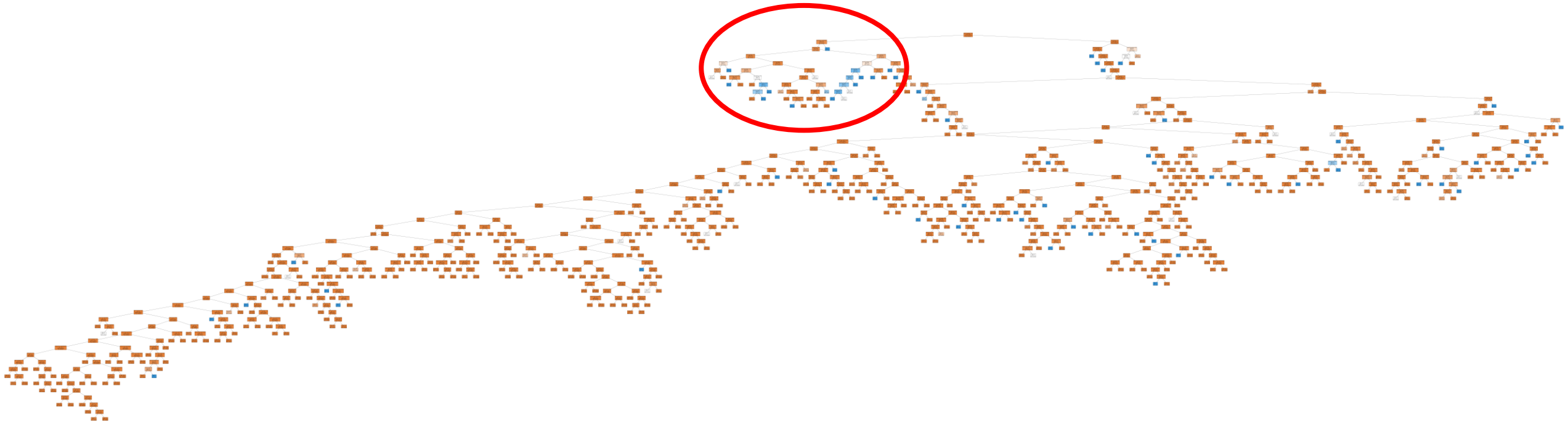
WICHITA STATE UNIVERSITY

PROPOSED WORK

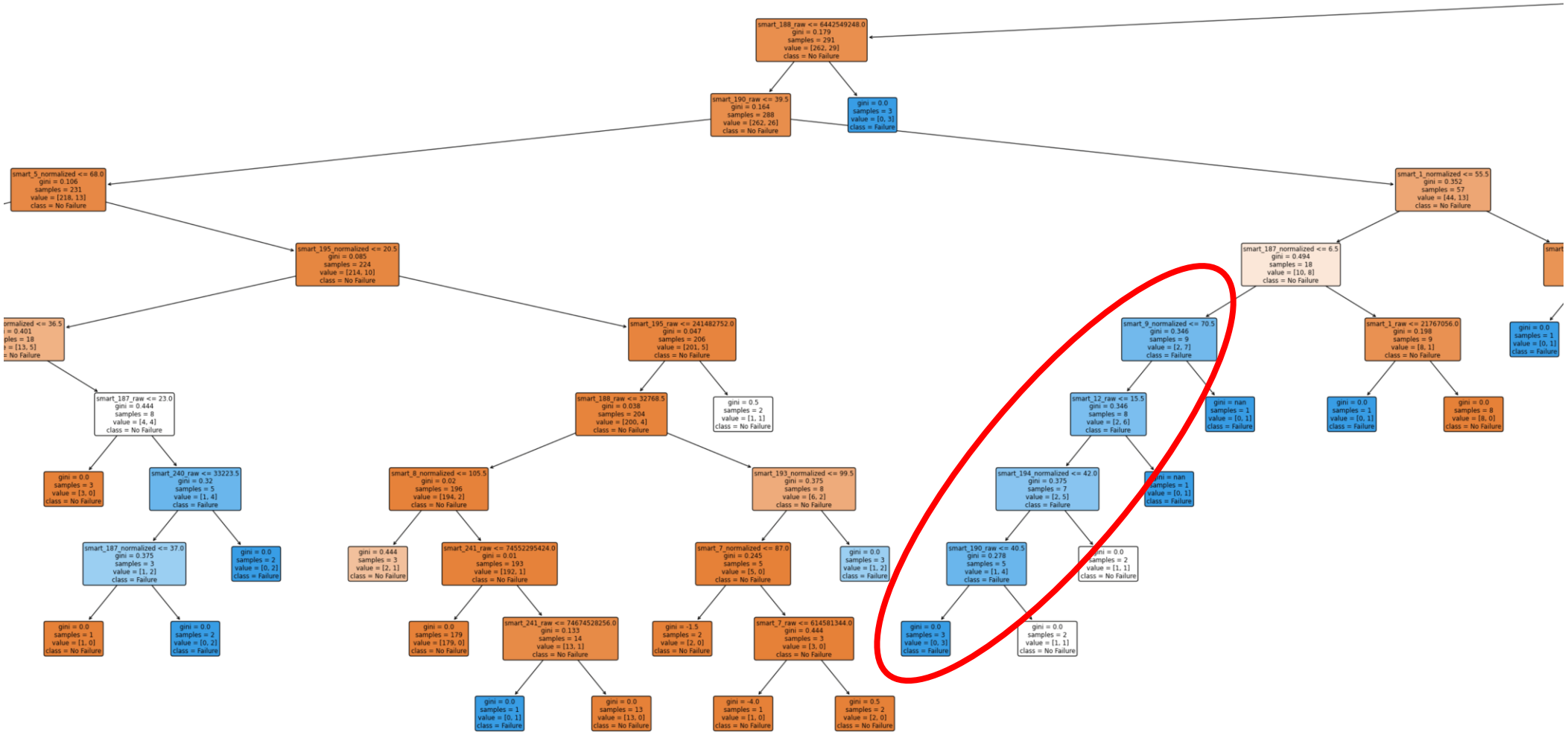
Methodology

- Legacy methods of determining a vintage of storage media had prior been a reactive action to a series of statistical analysis.
- New trends of Machine Learning have increased the potential of detecting these undesirable traits through analysis of Self-Monitoring, Analysis, and Reporting Technology “SMART” Logging.
- Source of our Training, Validating and Testing data will be Backblaze’s published Drive Stats.

DECISION TREE EVALUATION



DECISION TREE BRANCHES EVALUATION



APPLICATION

```
# %%
# Import libraries
import glob
import pandas as pd
import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import LSTM, Dense
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.preprocessing import MinMaxScalar
from sklearn.tree import DecisionTreeClassifier
from sklearn.metrics import accuracy_score
import matplotlib.pyplot as plt
# %%
# Load data
#df = pd.read_csv('/Users/scowley/Library/CloudStorage/OneDrive-WichitaStateUniversity/Coding/Python/intermediateConcepts/ECE777AB/20200701.csv')
# Get a list of all CSV files in the directory
files = glob.glob('/Users/scowley/Library/CloudStorage/OneDrive-WichitaStateUniversity/Coding/Python/intermediateConcepts/ECE777AB/Data/*.csv')
# %%
# Read each file into a DataFrame and store the DataFrames in a list
dfs = [pd.read_csv(file) for file in files]
# Concatenate the DataFrames
df = pd.concat(dfs)
# %%
# Select features and target
features = df.drop(['date', 'serial_number', 'model', 'capacity_bytes', 'failure'], axis=1)
target = df['failure']
# %%
# Normalize the features
scaler = MinMaxScaler(feature_range=(0, 1))
scaled_features = scaler.fit_transform(features)
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(scaled_features, target, test_size=0.2, random_state=42)
# Reshape input to be 3D [samples, timesteps, features]
X_train = X_train.reshape((X_train.shape[0], 1, X_train.shape[1]))
# %%
# Define the model
model = Sequential()
model.add(LSTM(50, activation='relu', input_shape=(X_train.shape[1], X_train.shape[2])))
model.add(Dense(1))
# Compile the model
model.compile(optimizer='adam', loss='binary_crossentropy', metrics=['accuracy'])

# %%
# Train the model
model.fit(X_train, y_train, epochs=200, verbose=0)
# Evaluate the model
X_test = X_test.reshape((X_test.shape[0], 1, X_test.shape[1]))
loss, accuracy = model.evaluate(X_test, y_test, verbose=0)
print('Test loss:', loss)
print('Test accuracy:', accuracy)
```

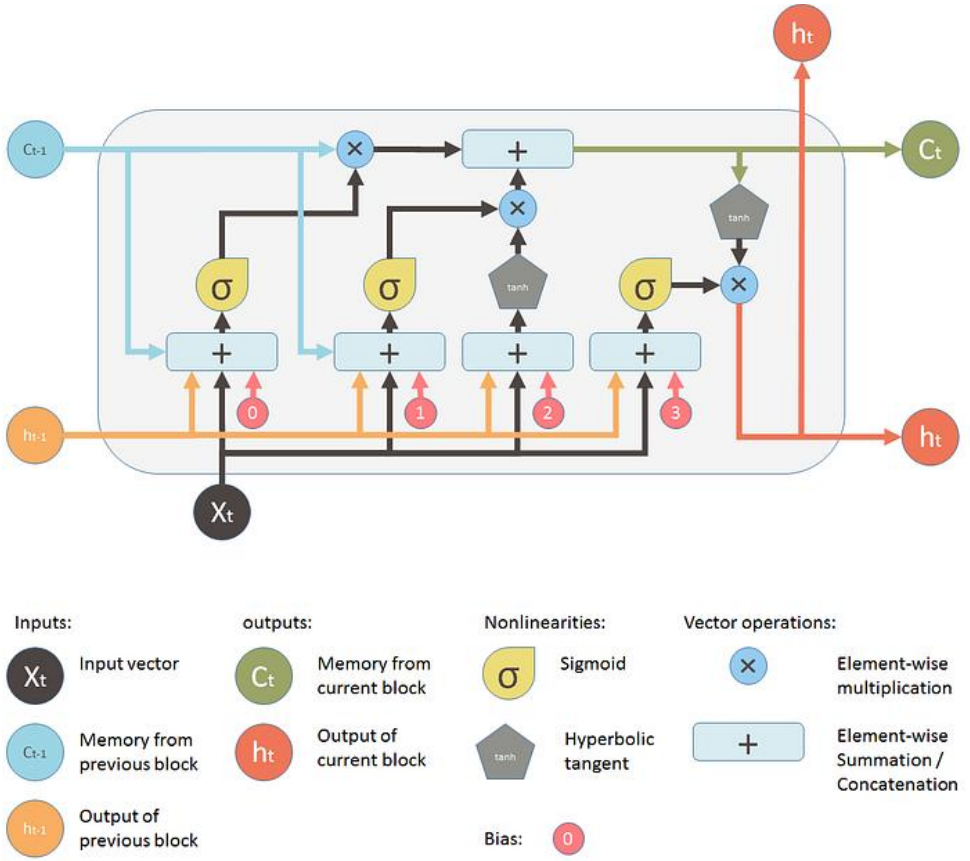
LSTM Layer with 50 Units
Dense Layer with 1 Unit

0 Hidden Layers declared, but many within the LSTM layer can be loosely defined as “Hidden”

WICHITA STATE UNIVERSITY

EVALUATION – LEARNINGS

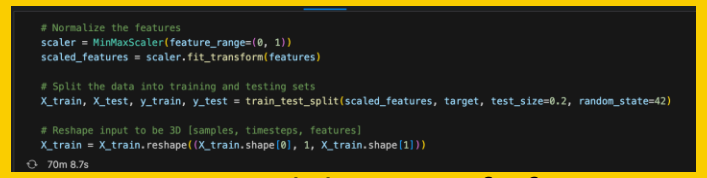
- Long Short-Term Memory (LSTM) is a Recurrent Neural Network (RNN) that utilizes feedback connections instead of feedforward Neural Networks.
  - An important feature of LSTM is the ability to 'remember' and retrieve information ove long periods of time, which is required for accommodating sequence data like ours.
  - **Operational perspective**
  - Forget Gate
    - Decides what information will be dropped from cell state
  - Input Gate
    - Updates cell state with new information
  - Update of Cell State
    - Actual cell state update
  - Output Gate
    - Decides what the next hidden state should be
- Data size; Overwhelmed our system with lack of failing results
  - Each CSV is a single day's worth of SMART Logs
    - 131 Columns (Attribute: SMART/SN/PN/Etc.), 150,000 Rows (Drives)
  - Each Quarter has 91 days of CSV Files
    - 13,650,000 Data Entries
  - Each Year has Four Quarters, 10 Years of Data
    - 546,000,000 Data Entries
  - Total Drive Count: 226,041
  - Total Failures: 12,722
  - Total POD: 333,011,602 Days
  - M1 Pro; 15 Minutes to model Decision Tree Classifier;
  - LSTM has significantly higher computational and time complexities.



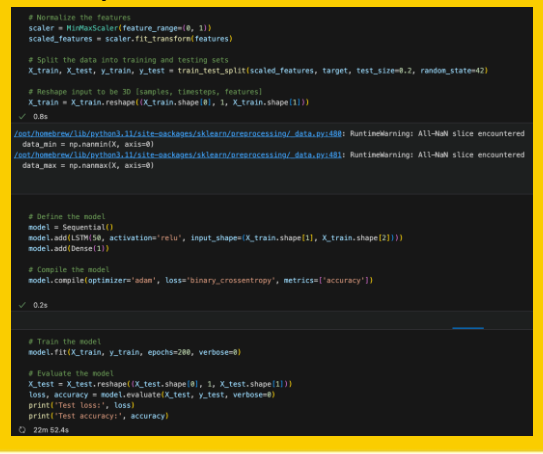
WICHITA STATE UNIVERSITY

CONCLUSIONS

- Failure to recognize system level limitations prevented full application of model.
- Rudimentary model was developed and proven by comparison to decision tree, but not diverse enough to be of real-world usage.
- 1Q used in this application has significant resource overhead which are constrained by system time outs.
  - IE: reshaping the training dataset to a 3D Object takes >45min; LSTM Model application is expected to take >2hr.



- 3 days DataSet to Validate Proof of Concept Suffers from the same issue. Large data sets are not good for Laptops.



WICHITA STATE UNIVERSITY

REFERENCES

- https://blog.mlreview.com/understanding-lstm-and-its-diagrams-37e2f46f1714
- https://www.backblaze.com/cloud-storage/resources/hard-drive-test-data
- https://scikit-learn.org/stable/tutorial/index.html
- https://www.tensorflow.org/tutorials
- Y. Y. Hu, S. Yoshida- S. Nakamura, K. Watanabe, W. Z. Lin, E. T. Ong, and J. Q. Mou, "Analysis of built-in speaker-induced structural-acoustic vibration of hard disk drives in Notebook PCs" IEEE Trans. Magn. 2009.
- J. Q. Mou, F. Lai, I. B. L. See and W. Z. Lin, "Analysis of structurally transmitted vibration of HDD in notebook computer," 2012 Digest APMRC, Singapore,
- J. Zhao et al., "Disk Failure Early Warning Based on the Characteristics of Customized SMART", 2020 19th IEEE Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems (ITherm), 2020.
- Z. Miller, O. Medaiyese, M. Ravi, A. Beatty and F. Lin, "Hard Disk Drive Failure Analysis and Prediction: An Industry View," 2023 53rd Annual IEEE/IFIP International Conference on Dependable Systems and Networks - Supplemental Volume (DSN-S), Porto Portugal, 2023

WICHITA STATE UNIVERSITY

Q/A & CONTACTS

- Sean Cowley
  - V688J944@Wichita.edu
  - www.linkedin.com/in/seancowley
- Yoel Woldeyes
  - J479Q742@wichita.edu
  - www.linkedin.com/in/yoel-sahle
- Md Reza E Rabbi
  - R844H794@wichita.edu
  - linkedin.com/in/md-reza-e-rabbi

WICHITA STATE UNIVERSITY