A comparative study of Neural network vs. Tree-based deep learning methods in the image classification of colorectal medical imagery diagnosis using HPCC supercomputing platform

Sarvesh Prabhu

Junior at Lambert High School, Graduating in May 2023

SarveshTamilPrabhu@gmail.com, Suwanee, GA, USA – 30024, (678) 822-2189

Mentors: Robert Foreman, Roger Dev, HPCC Systems, LexisNexis Risk Solutions Group, Alpharetta, GA

Abstract

The unprecedented breakthrough in the clinical trials for chronic illness diagnosis is a non-invasive SmartPill technology. One such chronic illness is colorectal cancer (aka colon cancer, CRC), third in prevalence and mortality among cancers in the United States. This internship furnishes an in-depth perspective on a DeepLearning-based CRC diagnosis and prognosis using the industry-standard HPCC platform leveraging ECL-ML and Generalized Neural Network Bundle (GNN).

This project is aimed to compare the performance of two well-established deep learning algorithms for medical imagery analysis, benefiting colorectal cancer diagnosis and prognosis. The project will compare two deep learnings methods: Neural network and Tree-based - Random Forest, Gradient Boosted, XGBoost, etc.

This solution involves four stages:

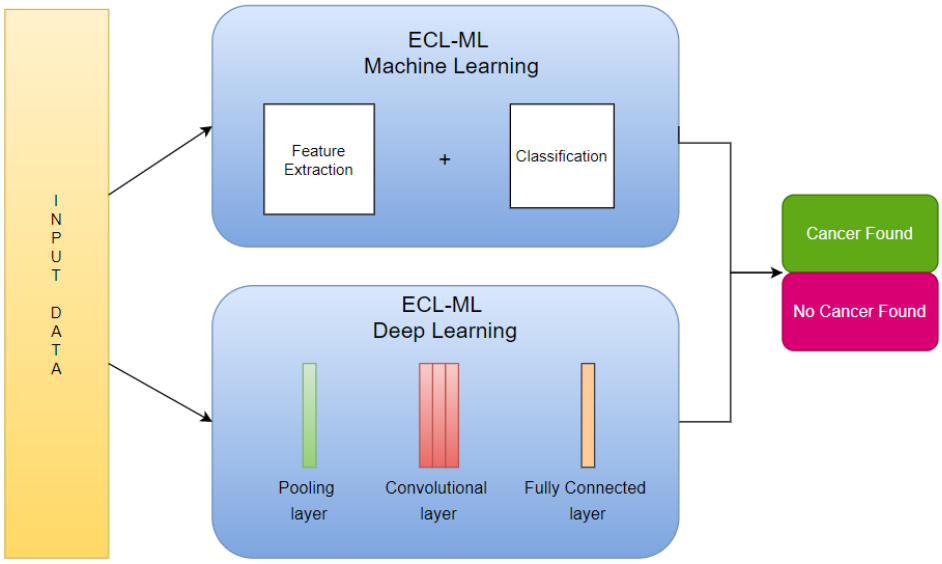
- **Stage 1 – Feature Extraction:** Standardizing images acquired by the SmartPill, perform classification and feature extraction utilizing the GNN bundle with the help of pre-defined annotations and labels.
- **Stage 2 – Convolutional Neural Network-based Deep Learning for Diagnosis:** By leveraging the ECL CNN library and utilizing Keras/TensorFlow to train a CNN model with 2 fully-connected layers harnessing GPU powered HPCC Thor.
- **Stage 3 – Tree-based Deep Learning for Diagnosis:** Use of convolutional autoencoders from stage-2 to extract image features with final classification by a tree-based model (Random Forest, Gradient Boosted Forest, etc.).
- **Stage 4 – Comparative Study & Conclusion:** Conduct a series of tests supporting a myriad of hypotheses, compare the model performance, efficacy of utilizing a tree-based model, complexity of hyper-tuning, the danger of overfitting, accuracy of prediction on a new dataset, implementation, and maintenance.

I am galvanized by the prospect of discovering the most pragmatic approach to medical imagery analysis by comparing the CNN and Tree-based learning outcomes. I am confident that this research will help the future of healthcare oncology practices, empowering medical practitioners to detect the undetected, and benefit patients by alleviating invasive procedures when they are already in chronic pain.

Deliverables

The deliverables include the fully working model utilizing the HPCC platform, annotation techniques, feature extraction methods, deidentified feature extracted data, ECL attributes, ECL models (incl. CNN, ECL-ML libraries), training dataset, coefficients, model, solution documentation, and research publication.

1. Data
    a. Source of raw data
    b. Image annotation schema
    c. Bias elimination documentation
    d. ECL-code for pre-processing and standardization of the data
2. Predictive Attributes
    a. Image classification
    b. Feature extraction using ECL-ML GNN bundle
    c. ECL attributes
3. Convolutional Neural Network and Tree-based Models
    a. Training dataset
    b. Model weights, Hyper-tuning
    c. ECL model development, Training, Hypothesis testing (p-value), error correction schemes
    d. Visualize model outputs
4. Comparative study & Research publication
    a. Comparison strategies
    b. Overall solution documentation



Timeline

1. Working with the Data & Configure HPCC Cluster — May-30 through Jun-10
    a. Procurement of CRC image dataset(s).
    b. Image annotation & labeling for the polyploids
    c. Review labeling with Dr. Saravanan, Pathanjali, MD
    d. Configure HPCC Thor cluster & Determine node size
2. Analytics Attributes & Training Data — Jun-13 through Jul-8
    a. Image classification, Feature extraction using GNN bundle
    b. Vessel structural or textural analysis of the image patches
    c. Hyper-tuning attributes
    d. Training data preparation
    e. Bias removal
3. Convolution Neural Network - Deep Learning Model — Jul-11 through Aug 5
    a. Hyper-tuning attribute: Diagnosis- Imagery processing based on CNN model
    b. Predictive model: Prognosis- Detection of tumor, isochromatic & diminutive polyps

```
# The following code is generated using Jupyter notebook using Python for
illustration purposes. I will be using the GNN bundle from the ECL-ML
library for the project implementation during the internship

from tensorflow.keras import layers

from tensorflow.keras import Model


# The input feature map of 125x125 with three color channels: R, G, and B

img_input = layers.Input(shape=(target_width, target_height,
color_channels))


# First convolution extracts 16 filters

# Convolution is followed by max-pooling layer

conX = layers.Conv2D(kernel_size, conv_window,
activation=activation)(img_input)

conX = layers.MaxPooling2D(pooling_window)(conX)


# Flatten feature map to a 1-dim tensor so we can add fully connected
layers

conX = layers.Flatten()(conX)
```

```
# Create a fully connected layer with ReLU activation and 512 hidden units
conX = layers.Dense(512, activation=activation)(conX)

# Add Droptout Regularization
conX = layers.Dropout(dropout)(conX)

# Create output layer with a single node and sigmoid activation
output = layers.Dense(len(class_names), activation = 'softmax') (conX)

# Create model: input = input feature map
# output = input feature map + stacked convolution/maxpooling layers +
fully connected layer + sigmoid output layer
model = Model(img_input, output)
```

(the green boxes show polyps detected by the algorithm under various lighting conditions)

c. Tree-based Model: Random Forest, Gradient Boosted, or XGBoost
d. Model training
e. Hypothesis testing (H0, H1, p-value)
f. Error correction (Type I and II)
g. Accept/Select region

4. Documentation & Visualization — Aug-8 through Aug-12
5. Comparative Study & Research publication — Aug-15 through Aug-26
6. Open-Source - Contribute artifacts to the HPCC community — Aug-29 through Aug-31

** On travel to Nashville, TN between 6/22-24, 2022 to participate in the HOSA international-level competition. It is accommodated in the timeline above.*

Wishlist

1. Implement the inference in the HPCC Roxie cluster - n/a -
2. Develop HL7 interface for EMR integration - n/a -
3. Integrate the model inference into NextGen EMR through HL7 - n/a -

Knowledge and Skills

Programming Language: Java 8

ML Libraries: Apache ML, Spark ML, WEKA 3, WEKA Workbench

Big Data: HPCC Platform ECL Introductory knowledge

Cloud: AWS - Proficient, Microsoft Azure - Introductory knowledge

IDE/Notebooks: VSCode, Jupyter, AWS SageMaker, and AWS Athena

Conclusion

By harnessing the power of the HPCC platform and the plethora of HPCC ECL-ML/Neural Networks/GNN libraries, I am confident that the model will predict the presence of Colorectal Cancer or polyps subject to the images provided. The model features the option of running real-time via Roxie or as a batch processing via Thor.

Both CNN and tree-based models are well-established in the data analytics industry and achieve high performance. The comprehensive comparative study will include the prediction accuracy on new images, statistical differences on the outcome, complexity of hyper-tuning, time to train, ability to do prognosis, maintainability, and whether the predictions are explainable.

The model receives the input images directly from a non-invasive SmartPill. With the help of Roxie, the model will compute in just a few hundred milliseconds and share the diagnosis with the doctor.

For the greater good, the image recognition capability and the model can extend to diagnose and the prognosis of diverticulitis, Colitis, Crohn's disease, and internal hemorrhoids in the future.

Additionally, the model can scale to support identifying chronic diseases in pets, making the SmartPill and diagnosis friendly for our fur family.

Thank you for the opportunity to present this proposal for your review and your consideration.

Sarvesh Prabhu

March 17, 2022