Turkish Journal of Computer and Mathematics Education *Vol.11 No.03 (2020), 1730-1734*
Research Article

Predicting Text using Machine Learning based on users unstrutured data

1st Zaheeruddin Ahmed
Computer Science
Charan Singh University

2nd Dr.B.L.Raju,
Professor & Principal Ahmed
Charan Singh University

3rd Dr. .Jatin Sharma,
Charan Singh University

***Abstract*— Many efforts are being made towards application of Machine Learning (ML) to solve Text Extraction problems, Text predictions to improve for betterment of society. In Educational institutions text extraction is widely practiced technique to examine inner elements of a document for clinical analysis and verification intervention, as well as for visual recognition of functioning of internal documents. This research project involves analysis of text from images for predicting text from documents available publicly. The research is subdivided into two major tasks, one is to build a model to understand usefulness of features in dataset, second to build a ML model for, then final task is to evaluate features extracted from processing document images for verification.**

Keywords*— Machine Learning, Text Extraction, image processing

I. INTRODUCTION

Machine Learning (ML) Machine learning is a subfield of artificial intelligence, which is broadly defined as the capability of a machine to imitate intelligent human behavior. Artificial intelligence systems are used to perform complex tasks in a way that is similar to how humans solve problems [1]. With the growing ubiquity of machine learning, everyone in business is likely to encounter it and will need some working knowledge about this field. A 2020 Deloitte survey found that 67% of companies are using machine learning, and 97% are using or planning to use it in the next year[2].

From manufacturing to retail and banking to bakeries, even legacy companies are using machine learning to unlock new value or boost efficiency. "Machine learning is changing, or will change, every industry, and leaders need to understand the basic principles, the potential, and the limitations," said MIT computer science professor Aleksander Madry, director of the MIT Center for Deployable Machine Learning.

The function of a machine learning system can be descriptive, meaning that the system uses the data to explain what happened; predictive, meaning the system uses the data to predict what will happen; or prescriptive, meaning the system will use the data to make suggestions about what action to take," the researchers wrote. There are three subcategories of machine learning:

Supervised machine learning models are trained with labeled data sets, which allow the models to learn and grow more accurate over time. For example, an algorithm would be trained with pictures of dogs and other things, all labeled by humans, and the machine would learn ways to identify pictures of dogs on its own. Supervised machine learning is the most common type used today.

In unsupervised machine learning, a program looks for patterns in unlabeled data. Unsupervised machine learning can find patterns or trends that people aren't explicitly looking for. For example, an unsupervised machine learning program could look through online sales data and identify different types of clients making purchases.

Reinforcement machine learning trains machines through trial and error to take the best action by establishing a reward system. Reinforcement learning can train models to play games or train autonomous vehicles to drive by telling the machine when it made the right decisions, which helps it learn over time what actions it should take.

Natural language processing

Natural language processing is a field of machine learning in which machines learn to understand natural language as spoken and written by humans, instead of the data and numbers normally used to program computers. This allows machines to recognize language, understand it, and respond to it, as well as create new text and translate between languages. Natural language processing enables familiar technology like chatbots and digital assistants like Siri or Alexa.

Recommendation algorithms. The recommendation engines behind Netflix and YouTube suggestions, what information appears on your Facebook feed, and product recommendations are fueled by machine learning. "[The algorithms] are trying to learn our preferences," Madry said. "They want to learn, like on Twitter, what tweets we want them to show us, on Facebook, what ads to display, what posts or liked content to share with us."

Image analysis and object detection. Machine learning can analyze images for different information, like learning to identify people and tell them apart — though facial recognition algorithms are controversial. Business uses for

this vary. Shulman noted that hedge funds famously use machine learning to analyze the number of cars in parking lots, which helps them learn how companies are performing and make good bets.

Fraud detection. Machines can analyze patterns, like how someone normally spends or where they normally shop, to identify potentially fraudulent credit card transactions, log-in attempts, or spam emails.

Automatic helplines or chatbots. Many companies are deploying online chatbots, in which customers or clients don't speak to humans, but instead interact with a machine. These algorithms use machine learning and natural language processing, with the bots learning from records of past conversations to come up with appropriate responses.

Self-driving cars. Much of the technology behind self-driving cars is based on machine learning, deep learning in particular. Medical imaging and diagnostics. Machine learning programs can be trained to examine medical images or other information and look for certain markers of illness, like a tool that can predict cancer risk based on a mammogram.

II. RESEARCH WORK

A. Objectives:

The objective of this research is the ability to search and find opensource data, define problem statement. Understand and learn image processing and segmentation. Develop machine learning models for extracted features. Evaluate predictions from different approaches and algorithms. In addition to this the aim was to develop analytical skills from real world data. Finally, this work is carried towards developing ML model to detect text from the set of phrases.

B. Technical Specifications:

To carry out this project we apply the model using the programming language Python with interface Anaconda and Jupiter lab, that include libraries /packages scikit-learn, OpenCV V2, TensorFlow 2.0, Karas, Pydicom, pandas, and NumPy on Python 3.8

C. Machine Learning approach

Machine learning algorithms used in the project work are decision tree, this is a decision support tool that uses a tree-like model of decisions and their possible consequences, including chance event outcomes, resource costs, and utility. Random forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean/average prediction (regression). Bootstrap aggregating or bagging is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. SVM support-vector machines are supervised learning models with associated learning algorithms that analyze data for classification and regression KNN The k-nearest neighbors (KNN) algorithm is a simple, easy-to-implement supervised machine learning algorithm that can be used to solve both classification and regression problems [4].

Logistic regression is a statistical model that in its basic form uses a logistic function to model a categorical variable. Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output. Stacking is an ensemble machine learning algorithm. It uses a meta-learning algorithm to learn how to best combine the predictions from two or more base machine learning algorithms.

C. Deep Learning approach

Deep learning (also known as deep structured learning) is part of a broader family of machine learning methods based on artificial neural networks with representation learning. Learning can be supervised, semi-supervised or unsupervised. Convolutional neural network (CNN) is a class of deep neural networks, most commonly applied to analyzing digital images. They typically contain convolution layers pooling, layers for processing image and reduce dimensionality. fully connected neural network layers and an output layer.

III. METHODOLOGY

Block diagram describes overall flow of the project work

Fig. 1. Research workflow

This project is further divided into different subtasks:

1. Developing Machine learning models for features in CBIS-DDSM data.
2. Process text extraction images and develop deep learning algorithms to predict cancer type.
3. Build Machine learning models for features extracted from images.

A. Information Gathering

Data used in the project are curated deidentified and available as opensource for academic work. Collection of feature data in CSV format. CBIS-DDSM is a database of 2,620 records. It contains normal text with verified pathology information and labels and features.. The CBIS-DDSM collection includes a subset of the DDSM data selected and

curated by a trained specialist. The images have been decompressed and converted to DICOM format.

B. Project Develeopment Life Cycle

After collecting data from different sources, it was loaded to Jupiter environment preprocessed and statistical analysis was done.

- CBIS-DDSM feature label data
- Data was loaded to pandas dataframe.
- Basic cleaning and EDA was performed to understand the basic nature and statistics of data.
- Categorical columns were encoded using LabelEncoder() for test and train data
- The merged test and train data was fed into various machine learning algorithms mentioned in Technology section earlier and did a 5-fold cross validation to select suitable algorithm for the data set.
- From cross validation better performing algorithms were selected and trained by fitting train data. Performed a hyperparameter tuning to get good fit parameters for final model.
- Final model performance was evaluated based on accuracy, precision, recall, f1 score for correct prediction.

B. Machine Earning for Text features

Procedure followed to evaluate usefulness of features extracted from data.

1. Data was in json files and those files were read and created a csv file and one dataframe to be able to use it conveniently in the future.
2. Data was split into train and test 70:30 ratio
3. Steps from 2-6 were repeated on the current dataset.

C. Deep Learning Model for Predicton

Deep learning models to predict breast cancer from Ultrasound images. Following is the description of procedure followed in developing deep learning model for ultrasound images.

1. Image data was read using OpenCV and preprocessed to extract region of interest (ROI).
2. Whole image and binary mask are divided into test and train and stored in separate folders with labels as directory names. Code snippet provided in later section [6].
3. Keras deep learning library provides a suite of convolutional layers which can be applied to image data. Image data can be visualized as array with 3 color channels or 1 grayscale and pixels. A complete CNN neural network is defined as shown in figure below [7].

Fig. 2. Architecture of CNN neural network used in the study

Image data from predefined library structure are read as train and test set. CNN network is trained for 30 epochs with 10 images as batch. Trained model is used to predict the cancer type from images. Prediction results are evaluated using accuracy, precision recall and confusion matrix as done earlier.

D. Research Project Design

CBIS DDSM dataset: The dataset has 13 input columns and 1 output column. The output column has 3 categories namely, Malignant, benign and benign_without,callback. Therefore, this is a multiclass classification problem. Below figures 7 shows the description of raining and testing data [5].

Various models were validated for performance using 5-fold cross validation. In this technique data is randomly partitioned into 5 equal sized subsamples. A single subsample is retained as the validation data for testing the model, and the remaining 4 sets are used as training data. Training and validation are repeated 5 times and results are averaged. Results are shown in table I II & III.

TABLE I. RESULTS OF CROSS VALIDATION

ML Algorithm	**accuracy**	**f1 score**	**precision**
DT	0.72	0.72	0.73
RF	0.76	0.75	0.74
Bag	0.74	0.74	0.75
Stack	0.75	0.74	0.74
XGB	0.75	0.75	0.75
Vot	0.7	0.7	0.7

ing	6	4	5
Logistic	0.75	0.73	0.71
KNN	0.76	0.75	0.76

TABLE II. FINAL MODELS AND THEIR PERFORMANCE SCORES

Model /Metric	Accuracy	Precision	Recall	F1 Score	Specificity
RF	0.76	0.83	0.59	0.61	0.76
XGB	0.86	0.87	0.74	0.77	0.86

Based on CV5 results selected Random forests and XGB for further training with data. Results are as shown in the table II.

D. Ultrasound Image Dataset

The ultrasound images are processed in opencv2, and tumor region or ROI is extracted. The histogram shows distribution of pixel values. When there is an anomaly observed in images pixel distribution has 2 or more peaks.

Note: Data is split into test and train and saved in different directories.

TABLE III. CV5 RESULTS FOR ULTRASOUND IMAGE FEATURES

ML Algorithm	Accuracy	F1 Score	Precision	Recall
DT	0.74	0.74	0.75	0.75
RF	0.78	0.77	0.78	0.78
Bag	0.78	0.75	0.78	0.78
Stack	0.81	0.80	0.80	0.81
XGB	0.77	0.77	0.77	0.77
Voting	0.79	0.79	0.79	0.79
Logistic	0.79	0.78	0.78	0.79
KNN	0.79	0.78	0.78	0.79

Fold Cross Validation and Model Selection and Training. Stacking model and Random forests were selected for further training and to classify images.

IV. TESTING METHODS & PROCEDURES

Performance evaluation of Machine learning models. Below figures show model performance evaluation results.

Fig. 3. Show performance evaluation results for machine learning models developed for features extracted from ultrasound images.

Fig. 4. Show performance evaluation results for 3 different deep learning models developed in the current project.

V. CONCLUSION AND DISCUSSION

Except for given data other models did not perform well. Machine learning models the one on extracted ROI could perform better than the model applied on whole image data. However, model has significantly high number of false positives and false negatives as data size may be low. It is promising to see that extracting features from images not only reduces noise in data predictions become more accurate. The model performance can be improved by various modifications such as Image augmentation to increase data size for training, or gathering more and more data, looking for more cues or features from the images.

Data integration from multiple sources can yield better predictions. By increasing data size improvement prediction accuracy and sensitivity can be achieved. By measuring size location depth of word, text analysis not only aid in computer aided diagnosis also be able to find location and depth of the word.

ACKNOWLEDGMENT

The authors would like to acknowledge the valuable inputs given by the guide, colleagues at university. Further we are very grateful to the staff of IT department for providing the infrastructure details. the support of all with inputs and discussions have been valuable in preparing this research paper.

REFERENCES

[1] WHO 2022, Accessed on: June 02, 2022. [Online] available https://www.who.int/news-room/fact-sheets/detail/cancer

[2] Alejandro Rodríguez-Ruiz, Elizabeth Krupinski, Jan-Jurre Mordang, Kathy Schilling, Sylvia H. Heywang-Köbrunner, Ioannis Sechopoulos, Ritse M. Mann, Detection of Breast Cancer with

Mammography: Effect of an Artificial Intelligence Support System, 2019, Radiology, Vol. 290, No. 2.

[3] Rodriguez-Ruiz A, Lång K, Gubern-Merida A, Broeders M, Gennaro G, Clauser P, Helbich TH, Chevalier M, Tan T, Mertelmeier T, Wallis MG, Andersson I, Zackrisson S, Mann RM, Sechopoulos I. Stand-Alone Artificial Intelligence for Breast Cancer Detection in Mammography: Comparison With 101 Radiologists. J Natl Cancer Inst. 2019 Sep 1;111(9):916-922. doi: 10.1093/jnci/djy222. PMID: 30834436; PMCID: PMC6748773.

[4] McKinney, S.M., Sieniek, M., Godbole, V. et al. International evaluation of an AI system for breast cancer screening. Nature 577, 89–94 (2020). https://doi.org/10.1038/s41586-019-1799-6

[5] Aly Fahmy, OpenScholor, Accessed on: June 02, 2022. [Online] available, https://scholar.cu.edu.eg/?q=afahmy/pages/dataset

[6] Cancer Imagin Archive, Accessed on: June 02, 2022. [Online] available, https://wiki.cancerimagingarchive.net/display/Public/CBIS-DDSM

[7] Rebecca Sawyer Lee, Francisco Gimenez, Assaf Hoogi , Daniel Rubin (2016). Curated Breast Imaging Subset of DDSM [Dataset]. The Cancer Imaging Archive.

[8] Kaggle, Accessed on: June 02, 2022. [Online] available :Referred Kaggle/ Git / Stackoverl