Artificial Intelligence-Based Handwriting Recognition

Prepared for: DATS6202 Machine Learning I
Prepared by: Group 6, Reema Domadia, Dinesh Kumar Padmanabhan, Utkarsh Nigam

Introduction:

Sixty-one percent of medication errors in hospitals are a result of illegible handwriting and transcription errors, according to a sciencedaily.com article titled “Computerized Doctors' Orders Reduce Medication Errors”. (Center for the Advancement of Health, 2007) Historically, the notes that doctors scribbled into a patient's medical history record file were generally only seen by doctors. However now, doctors are just one component of a multidisciplinary healthcare network. A consequence of this is that illegible records composed by rushed doctors are now presented to colleagues with no qualifications in cryptology. In an article by The National Medical Journal of India, the author, Santosh Chaturvedi, unveils an alarming statistic: 7,000 patient deaths a year as a result of illegible handwriting. (Chaturvedi, 2018) By using an AI-driven handwriting recognition software doctors can continuously work undisruptioned with the software working to convert the handwriting into a computerized legible script. The advantage of this software is that the script can change multiple hands and everyone will receive the same message, leaving no room for misinterpretation. This capability is crucial for industries like healthcare as medical records are often passed onto a web of professionals/specialists, ranging from doctors to pharmacists, to complete the care of a patient. A minor miscommunication due to illegible handwriting care causes the patient their life.

Problem Statement and Dataset:

Due to this market need, the purpose of this research project is to bridge a known gap in medical communication and record keeping by training a model that will identify and decrypt handwriting to convert into a computer-typed text through a neural network. While complete accuracy will be achieved over-time by training the model with a larger data sample, to achieve the scope of this prototype we will be using a dataset created by the National Institute of Standards and Technology (NIST). The dataset, as it was found, does not require any cleaning and is equipped to fuel the development of this project. The NIST Special Database 19 consists of 800,000 handwriting sample form (png) images serving as observations in our dataset. Those characters each occupy 28x28 pixels per raster and are labeled by one of 62 ASCII hexadecimal classes corresponding to the letters “A”- “Z” . The abundance of data contained in this database will suffice to train our model. To be able to meet the deadlines for this project, the goal of the prototype will be able to identify and classify characters. Once the prototype is built, the stakeholders can further develop this project by implementing a loop that will identify and classify every character forming works, sentences, and further paragraphs.

Methodology

To promote collaboration in analyzing this problem, GitHub will serve as the platform to compile all efforts made by members of the team. We will utilize the PyCharm IDE with the assistance of Anaconda libraries, often referred to as Conda, to employ a combination of neural networking and machine learning techniques to obtain the desired output. Due to the versatility of PyCharm and Conda libraries, we anticipate that the sole software is expected to suffice the technical needs of this project with notepad serving as the user input application. Respectively, by leveraging the power of a neural network, a multilayer perceptron with backpropagation enables the model to analyze each handwritten letter to which then the machine learning technique, classification, will group it with the appropriate lettering or numerical category. Since this model is a prototype, this particular model will be able to decrypt handwriting and convert it into an easy-to-read computer-generated lettering format to which stakeholders will be able to further develop by implementing loops that will decrypt words, phrases, sentences and further, paragraphs.

Evaluating Performance:

To obtain domain knowledge, we will be looking at the research conducted by The National Institute of Health, all housed in the US National Library of Medicine's Database. Specifically, an article titled "Poor handwriting remains a significant problem in medicine" by Daniel K Sokol and Samantha Hettige, has created a need for such a capability and laid the foundation for this project. Upon obtaining a sufficient understanding of the problem, the team expects to implement necessary machine learning and neural networking algorithms as taught by Professor Amir Jafari in the Machine Learning I and Introduction to Data Mining courses. The performance of the network will be generated by training and testing the model with portions of the dataset. To assess the models validity, we will be employing a combination of a confusion matrix, sum squared error rate, and mean error rate. Through these assessments we will be able to identify and mitigate any misclassifications of characters ensuring optimal model performance. Recognizing errors can not only mitigate potential biases but also identify the associated type I and type II errors. By doing so, we will enable the model to output optimal results that will enhance the users' recordkeeping and communication capability. This coupled with pre-existing domain research on this topic by the NIH, we will be able to bridge a gap in the market needs.

Schedule:

To ensure our project is on track we will pursue with below schedule (Figure 1).

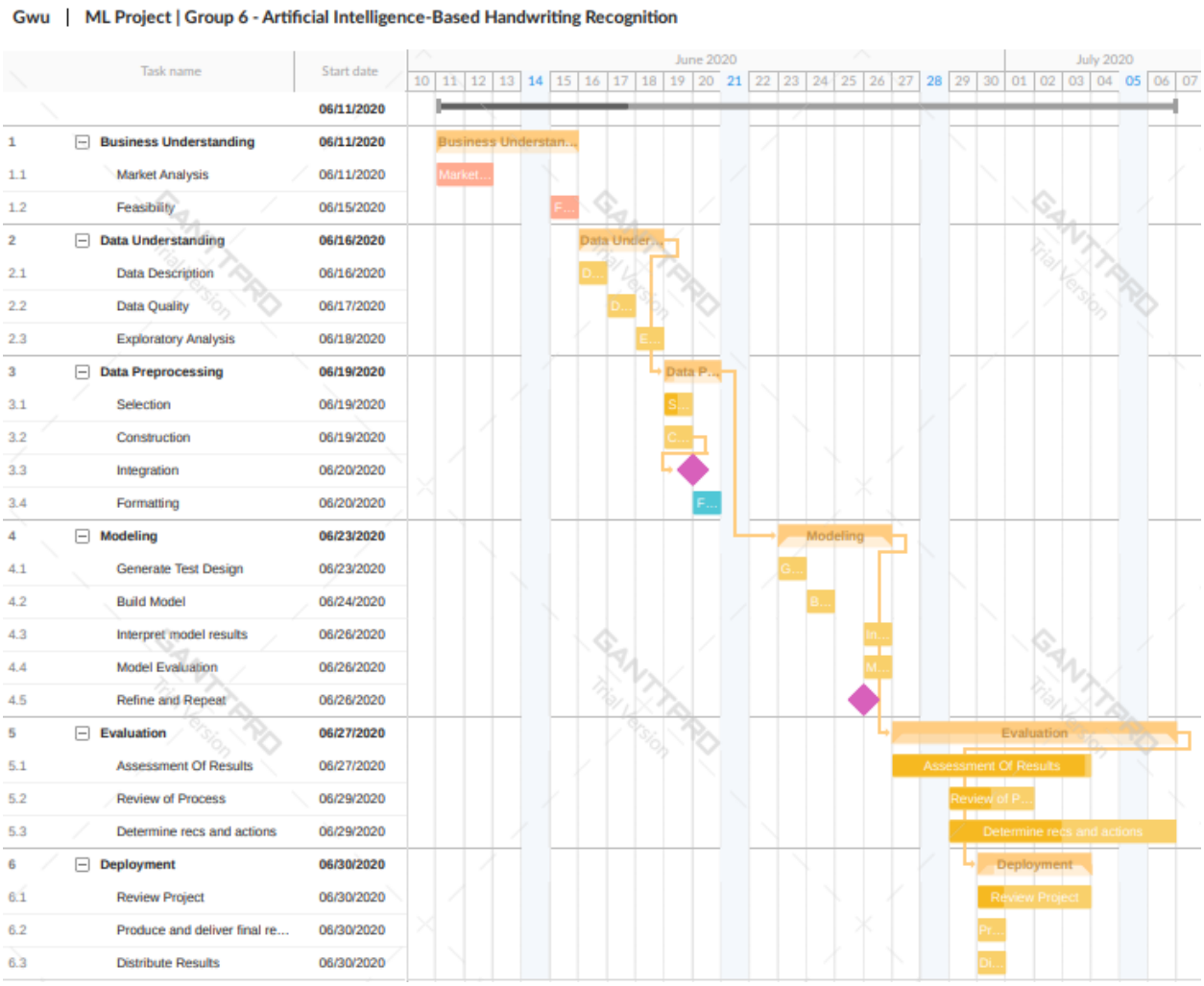

Figure 1: Schedule for Project

Reference:

Center for the Advancement of Health. "Computerized Doctors' Orders Reduce Medication Errors." ScienceDaily.ScienceDaily,27June2007. <www.sciencedaily.com/releases/2007/06/070627084702.htm>.

Chaturvedi SK. What's wrong with doctors' handwriting?. Natl Med J India [serial online] 2018 [cited 2020 Jun 18];31:47-8. Available from: < http://www.nmji.in/text.asp?2018/31/1/47/243417 >.

Grother, Patrick J. "NIST Special Database 19." NIST, National Institute of Standards and Technology, 27 Apr. 2019, < www.nist.gov/srd/nist-special-database-19 >.

Sokol, Daniel K, and Samantha Hettige. "Poor handwriting remains a significant problem in medicine." Journal of the Royal Society of Medicine vol. 99,12 (2006): 645-6. doi:10.1258/jrsm.99.12.645