A Web-based Application for the Prognostication of COVID-19 Patient Outcomes Through the Application of Machine Learning and Deep Learning Techniques on A Heterogeneous Dataset of CT-Chest Images & Clinical Data

A Capstone Technical Document
Submitted in Partial Fulfilment of the Requirement for the Degree of
Master of Science in Applied Data Science

of
The University of the West Indies

Elombe Calvert
2022

Department of Computing
Faculty of Science and Technology
Mona Campus

A Web-based Application for the Prognostication of COVID-19 Patient Outcomes Through the Application of Machine Learning and Deep Learning Techniques on A Heterogeneous Dataset of CT-Chest Images & Clinical Data

Elombe Calvert[1], Gunjan Mansingh[1], Ricardo Anderson[2]

ABSTRACT

Introduction: COVID-19 was named a pandemic on March 11, 2020, by the World Health Organization (WHO). Unfortunately, patients with COVID-19-associated acute-respiratory distress syndrome (ARDS) frequently require intubation and intensive care unit (ICU) care, both of which are costly and limited, especially within developing countries. Thus, knowing a patient's prognosis shortly after admission to the hospital can help with starting new medicines and therapies, resulting in better patient outcomes. This, although difficult, can be done reliably by applying computational approaches to the complex heterogeneous biomedical data obtained from COVID-19 hospitalized patients. With that said, this project aims to develop a clinical decision support system for physicians that will aid in the accurate determination of a COVID-19 patient's prognosis at admission, whether they will die or recover, based on clinical, biological, and radiological data collected. **Methods:** For the categorization of CT chest images into categories of normal lung, no lung tissue, and lung tissue with pathology, a modified VGG-16 was utilized, with 13 layers. This model architecture was also used to classify the top 10 positive patients' CT scans into recovered or deceased. A 12-layer multilayer perceptron was used to classify the structured data into recovered or deceased. The final prediction of a patient's prognosis was done using a logistic regression model utilizing ridge regression. **Results:** The most successful outcome was delivered by the 13-layer CNN trained on CT images, obtaining an accuracy of 91 percent, as compared to using the structured data with an accuracy of 70 percent and a final prediction of 58 percent using logistic regression and L2 regularization. **Conclusion:** Complex, numerous biomedical data from covid-19 positive patients can be used to train models with accurately predict if a patient will recover or die from their infection. These models can be deployed as a web-based application for use by healthcare professionals,

Index Terms – Machine Learning**,** Deep Learning, COVID-19, Data, Model, CT-scan, Prognosis, Pandemic

INTRODUCTION

SARS-CoV-2, a novel coronavirus, was discovered following an outbreak of unexplained viral pneumonia cases in the Wuhan region of mainland China in December of 2019. The World Health Organization (WHO) declared the outbreak a Public Health Emergency of International Concern on January 30, 2020, and a pandemic on March 11, 2020. Over the next 20 months, COVID-19 pneumonia, the name designated by the World Health Organization as a viral pneumonia caused by SARS-CoV-2, has affected millions of individuals and has taken countless lives globally (Zhu & Gallego, 2021).

COVID-19 causes a viral pneumonia that is responsible for respiratory problems that can rapidly progress to acute respiratory failure which has placed a considerable burden on healthcare systems around the world as a result of a large number of critically ill individuals hospitalized. As we have come to now know, COVID-19 patients with acute respiratory distress syndrome (ARDS) frequently require intubation and intensive care unit (ICU) care, both of which are costly and limited, especially within developing countries. Because mechanical ventilators and ICU care are in short supply, it is vital to precisely and quickly predicts COVID-19 patients who will progress to having advanced disease, thus allowing for the early initiation of aggressive interventions needed to help prevent unfavorable outcomes. Thus, knowing a patient's prognosis shortly after admission can help with starting new medicines like remdesivir, tocilizumab, convalescent plasma transfusion, and other developing therapeutics, resulting in better patient outcomes (Wang et. al., 2022). In COVID-19, all patients, especially those with comorbidities including obesity, cardiovascular disease, chronic lung disease, hypertension, or cancer, are at risk of developing critical illness after hospitalization. However, predicting when a COVID-19 patient would advance to critical illness remains difficult, even for experienced clinicians.

The truth is, for a patient who is admitted to the hospital for COVID-19-associated pneumonia, it can be a very daunting and challenging task for a physician to fully grasp and assimilate the different streams of heterogenous data collected from a patient and use the derived insights to accurately predict how a patient's condition will progress over time. It is without question that the biomedical data collected from a patient is complex, extensive, and heterogenous, with each parameter having its own inherent relationship to all the other data parameters collected. The typical assortment of data collected from an admitted COVID-19 patient consists of biographical and historical data, examination findings, vital signs, radiologic images, and biochemical and laboratory data. To completely assess the non-linear patterns that exist in these complex biomedical datasets, insights which are hidden from human intuition, the use of computational approaches and data science techniques can prove to be advantageous in improving a patient's prognosis.

The promise of applying artificial intelligence to biomedical data, thus prognosticating the outcome of COVID-19 patients is reasonable, as patient demographics, laboratory parameters, and CT-chest images have been shown to correlate well with the severity of pneumonia documented in persons infected with SARS-CoV-2. Furthermore, machine and deep learning approaches when utilized will allow for large amounts of biomedical data to be integrated and processed, thereby deriving meaningful insights regarding a patient's disease severity and the likelihood of progression to an adverse outcome. Thus, through this project, a reliable clinical decision support system for physicians will be created, allowing for a more efficient assessment of a COVID-19 patient's disease status at admission and more accurate predictions on how the patient's disease will evolve over the course of their admission. This ultimately will lead to patients with predicted poorer outcomes being started on vital treatments and interventions much sooner, leading to less critically ill patients, better utilization of hospital resources, and a healthcare ecosystem that is not overwhelmed by very sick patients.

METHODOLOGY

The main objective of this research project is to integrate clinical, biological, and radiological data using machine learning and deep learning models to predict the prognosis of hospitalized COVID-19 patients, by utilizing a web-based application platform.

Data Description

The Integrative CT Images and Clinical Features for COVID-19 (iCTCF) dataset were obtained from Kaggle; a website that hosts numerous open-source datasets for the purpose of research and the advancement of machine learning through competitions.

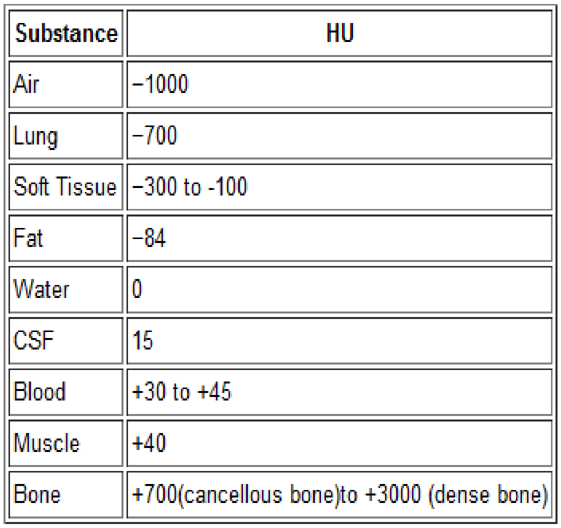
The iCTCF dataset was curated from two hospitals in Wuhan, China: Union Hospital and Liyuan Hospital. The iCTCF dataset includes data acquired from hospitalized patients over two time periods, November 14 to November 30, 2019, and January 25 to February 20, 2020 (Ning et. al., 2020). The data set originally contained 1,521 individuals, including 1,126 from Union Hospital and 395 from Liyuan Hospital. The dataset included 756 males and 765 females. 345 of the subjects had ages less than 40 years old, 531 were 40 to 60 years old and 645 patients were over the age of 60 years old.

The dataset comprises patients with laboratory-confirmed COVID-19 and those who are COVID-19-negative, which includes those suspected of having COVID-19 but have a negative test and non-COVID-19 patients introduced into the dataset as a part of a control group. 894 patients had a positive COVID-19 test, versus 627 who had a negative test or no test done. Of the 894 patients reported to be positive for COVID-19, 662 were documented to be cured, 57 deceased, and 175 whose mortality outcome was unknown, possibly due to transfer to another hospital within that healthcare ecosystem. It should be noted that the iCTCF dataset consists of two distinct categories of data: clinical features data and computer tomography chest (CT-chest) images. All patients included had clinical features data but only 1,342 subjects had both CT-chest images and clinical features data. The clinical features data from 1,521 individuals are classified into 130 types from 9 categories, including basic information, routine blood test, inflammation test, blood coagulation test, biochemical test, immune cell typing, cytokine profile test, autoimmune test, and routine urine test. There are 125 clinical features and 5 which are categorical, the shape of which includes 130 columns and 1,521 rows. On the other hand, CT-chest images amounted to 364,357 slices which were exported from 1,342 subjects with the CT-chest data in a DICOM format and then converted to jpeg (Ning et. al., 2020).

Data Preparation Techniques

The unstructured data used for this project are CT-chest images derived from 3D CT scans of a patient's thorax (lung fields). A CT scan is a greyscale image consisting of different areas of densities represented by shades of dark and white areas. These areas of varied densities measured in Hounsfield Units (HU) reflect different structures within a patient's chest as shown in table (1) below.

Table 1 Substances within the body and their associated Hounsfield Unit.



Substance	HU
Air	−1000
Lung	−700
Soft Tissue	−300 to -100
Fat	−84
Water	0
CSF	15
Blood	+30 to +45
Muscle	+40
Bone	+700(cancellous bone)to +3000 (dense bone)

By plotting the cumulative Hounsfield Units of the CT-chest images shown in figure (1) below, it was found that the images contained a lot of densities that did not reflect lung tissue, such as soft tissue, fat, water, and bone. Due to this revelation, the images were denoised or segmented using **Mean Adaptive Thresholding** using the **Sobel Operator** for edge detection. The resulting CT-Chest images after segmentation contained only areas of lung tissue.

Figure 1 Histogram of Hounsfield Unit distribution of CT-chest images used in the project.

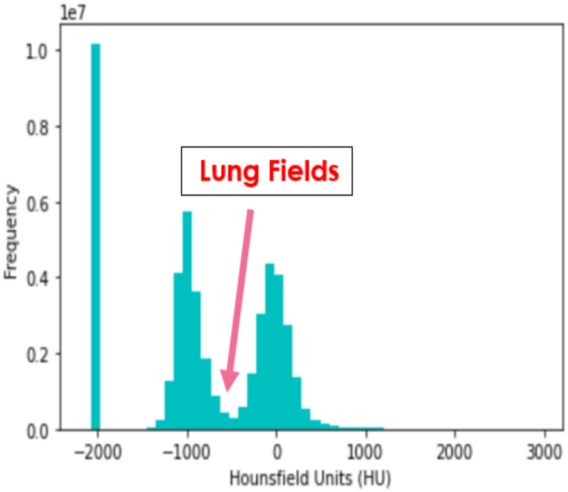


Figure 2 Unsegmented, Noisy CT-chest Images from an index patient from the dataset.

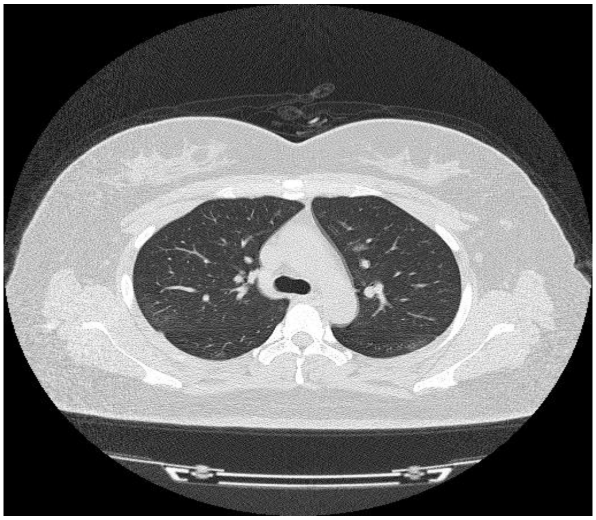
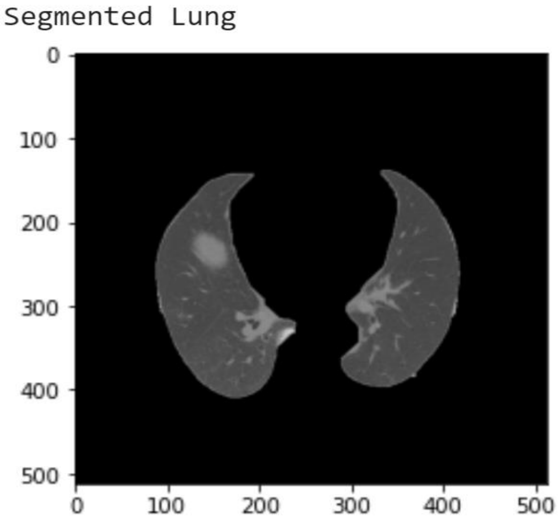
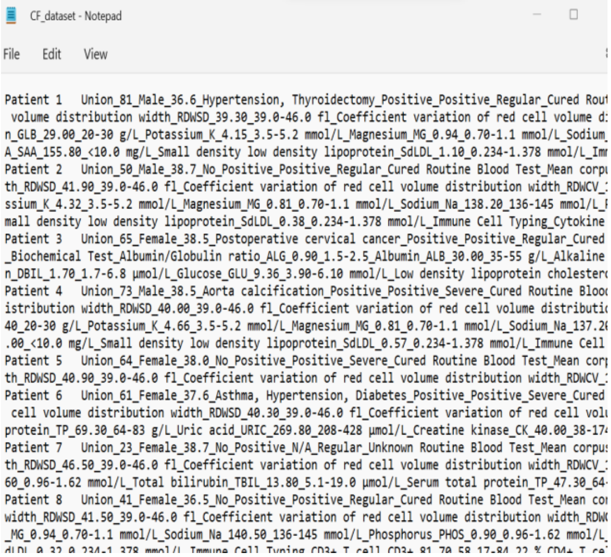


Figure 3 Segmented, Denoised CT-chest Images from an index patient from the dataset.



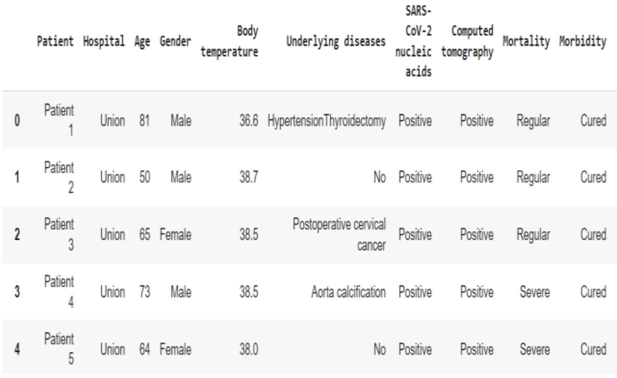
The structured data used for this project was extracted as a text file with the data in an unstructured form as shown in figure (4) below. Through the creation of a custom parser that employed regular expressions from the pandas library regex and string manipulation, a data frame of the data was created shown in figure (5) below.

Figure 4 Text file containing patient clinical data in an unstructured form.



```
Patient 1   Union_81_Male_36.6_Hypertension, Thyroidectomy_Positive_Positive_Regular_Cured Rout
 volume distribution width_RDWSD_39.30_39.0-46.0 fl_Coefficient variation of red cell volume di
n_GLB_29.00_20-30 g/L_Potassium_K_4.15_3.5-5.2 mmol/L_Magnesium_MG_0.94_0.70-1.1 mmol/L_Sodium_
A_SAA_155.80_<10.0 mg/L_Small density low density lipoprotein_SdLDL_1.10_0.234-1.378 mmol/L_Imm
Patient 2   Union_50_Male_38.7_No_Positive_Positive_Regular_Cured Routine Blood Test_Mean corpu
th_RDWSD_41.90_39.0-46.0 fl_Coefficient variation of red cell volume distribution width_RDWCV_
ssium_K_4.32_3.5-5.2 mmol/L_Magnesium_MG_0.81_0.70-1.1 mmol/L_Sodium_Na_138.20_136-145 mmol/L_
mall density low density lipoprotein_SdLDL_0.38_0.234-1.378 mmol/L_Immune Cell Typing_Cytokine
Patient 3   Union_65_Female_38.5_Postoperative cervical cancer_Positive_Positive_Regular_Cured
_Biochemical Test_Albumin/Globulin ratio_ALG_0.90_1.5-2.5_Albumin_ALB_30.00_35-55 g/L_Alkaline
n_DBIL_1.70_1.7-6.8 µmol/L_Glucose_GLU_9.36_3.90-6.10 mmol/L_Low density lipoprotein cholester
Patient 4   Union_73_Male_38.5_Aorta calcification_Positive_Positive_Severe_Cured Routine Bloo
istribution width_RDWSD_40.00_39.0-46.0 fl_Coefficient variation of red cell volume distributi
40_20-30 g/L_Potassium_K_4.66_3.5-5.2 mmol/L_Magnesium_MG_0.81_0.70-1.1 mmol/L_Sodium_Na_137.2
.00_<10.0 mg/L_Small density low density lipoprotein_SdLDL_0.57_0.234-1.378 mmol/L_Immune Cell
Patient 5   Union_64_Female_38.0_No_Positive_Positive_Severe_Cured Routine Blood Test_Mean cor
th_RDWSD_40.90_39.0-46.0 fl_Coefficient variation of red cell volume distribution width_RDWCV_
Patient 6   Union_61_Female_37.6_Asthma, Hypertension, Diabetes_Positive_Positive_Severe_Cured
 cell volume distribution width_RDWSD_40.30_39.0-46.0 fl_Coefficient variation of red cell vol
protein_TP_69.30_64-83 g/L_Uric acid_URIC_269.80_208-428 µmol/L_Creatine kinase_CK_40.00_38-17
Patient 7   Union_23_Female_38.7_No_Positive_N/A_Regular_Unknown Routine Blood Test_Mean corpus
th_RDWSD_46.50_39.0-46.0 fl_Coefficient variation of red cell volume distribution width_RDWCV_
60_0.96-1.62 mmol/L_Total bilirubin_TBIL_13.80_5.1-19.0 µmol/L_Serum total protein_TP_47.30_64
Patient 8   Union_41_Female_36.5_No_Positive_Positive_Regular_Cured Routine Blood Test_Mean co
width_RDWSD_41.50_39.0-46.0 fl_Coefficient variation of red cell volume distribution width_RDW
_MG_0.94_0.70-1.1 mmol/L_Sodium_Na_140.50_136-145 mmol/L_Phosphorus_PHOS_0.90_0.96-1.62 mmol/L_
dLDL_0.32_0.234-1.378 mmol/L_Immune Cell Typing_CD3+ T cell_CD3+_81.70_58.17-84.22 %_CD4+ T ce
```

Figure 5 Dataframe made from unstructured data in text file shown above.



	Patient	Hospital	Age	Gender	Body temperature	Underlying diseases	SARS-CoV-2 nucleic acids	Computed tomography	Mortality	Morbidity
0	Patient 1	Union	81	Male	36.6	HypertensionThyroidectomy	Positive	Positive	Regular	Cured
1	Patient 2	Union	50	Male	38.7	No	Positive	Positive	Regular	Cured
2	Patient 3	Union	65	Female	38.5	Postoperative cervical cancer	Positive	Positive	Regular	Cured
3	Patient 4	Union	73	Male	38.5	Aorta calcification	Positive	Positive	Severe	Cured
4	Patient 5	Union	64	Female	38.0	No	Positive	Positive	Severe	Cured

Fields within the structured dataset that represented laboratory tests administered to patients were entirely continuous variables. Discrete fields within the dataset were “Hospital”, “Gender”, “Underlying diseases”, “Morbidity”, etc. All continuous variables were normalized using min-max normalization, where the maximum value for a feature is transformed to 1 and the minimum to 0. Discrete variables were encoded into numerical categories. Where missing values were found for a continuous variable, it was replaced by 0.5.

Data Modelling

Both structured and unstructured data forms represented in this project were modeled using machine learning and deep learning methods, taking the form of convolutional neural networks, a normal multilayer perception, and a logistic regression model.

CT Scan Slice Categorization

For a typical patient who has done a CT scan of the chest, the slices/image segments that are acquired may include slices containing no lung tissue, lung tissue that is normal with no disease, and slices affected by the disease. In order to make this natural distinction, a convolutional neural network (CNN) was trained on available labeled CT image data in order to make such differentiations. The CNN used for this classification task was derived from the VGG-16 model. In order to decrease model complexity to enable better abstraction of image features, three convolutional layers were dropped from the model's architecture, turning it into a 13-layer network with inputs (200,200,1), six convolutional layers, three max-pooling layers, and two dense layers. This model architecture employed two drop-out layers at 0.5, learning rate 0.001, decay 0.05, batch size 64, at 300 epochs.

Prognosis Prediction: Utilization of CT Chest Images

In order to make predictions about prognosis, it is vital to select the best right images for training; this is done by selecting the 10 most probable CT chest images for each patient. These 10 images will be stacked horizontally into a 3D vector having 10 channels and passed as an input to the convolutional neural network for training. The convolutional neural network used for this binary classification problem for predicting morbidity (recovered or deceased) is an augmented VGG-16 with 13 convolutional layers instead of 16; as mentioned before, the input to this architecture is 200, 200,10. This model has six convolutional layers, three max-pooling layers, and 2 dense layers. Two drop-out layers at 0.5 were employed, with a learning rate of 0.0007, decay 0.05, and batch size of 64, at 300 epochs.

Prognosis Prediction: Utilization of Structured Data

The prepared structured data used for this research project consist of biomedical data for each patient, which includes personal data and results from different blood tests done upon admission. In using this form of data to predict prognosis, a 12-layered multilayer perceptron was used. The inputs to this feed-forward neural network were the 127 training features from the dataset. This network consisted of five dense layers, five dropout layers with values of 0.5 and 0.2, and one output layer. This architecture employed a learning rate of 0.0007, a decay of 0.05, batch size of 64, at 300 epochs.

Final Predictions: Combining Previous Probabilities

In order to make final predictions on the prognosis of a patient, the previous two prediction probabilities will be used to predict the final prognosis of a patient. To facilitate such as endeavor, logistic regression with ridge (L2) regularization is employed. This method works on the principle of introducing an additional parameter, called the penalty, which allows the model not to overfit the very small training dataset while optimizing the final prediction. This model outputs a value between 0 and 1, representing the possibility of dying or recovering from their covid-19 infection.

RESULTS

This section of the paper gives a detailed overview of how the different models were evaluated and the results obtained from the output of each architecture.

Evaluation: CNN CT SCAN CATEGORIZATION

In evaluating this 13-layer convolutional neural network, accuracy, loss, and AUC were evaluated for both training and validation. Training metrics; accuracy 0.9174, loss 0.3035, AUC 0.9794, validation metrics; accuracy 0.8095, loss 0.4633, AUC 0.9388.

Figure 6 Training and Validation Curves Shown for Accuracy Metric.

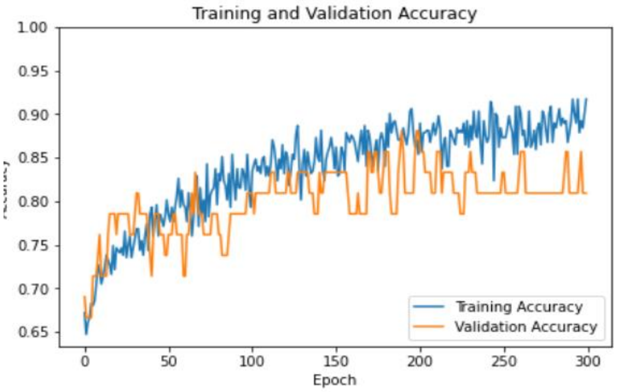


Figure 7 Training and Validation Curves Shown for Loss Metric.

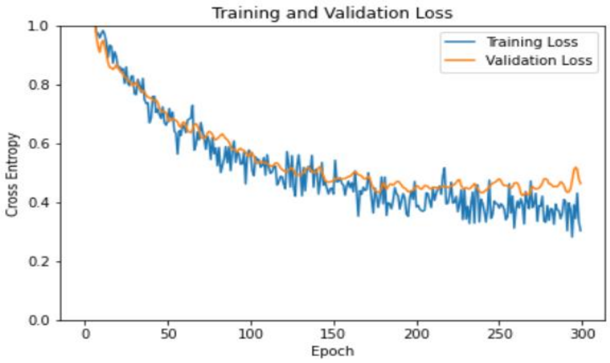
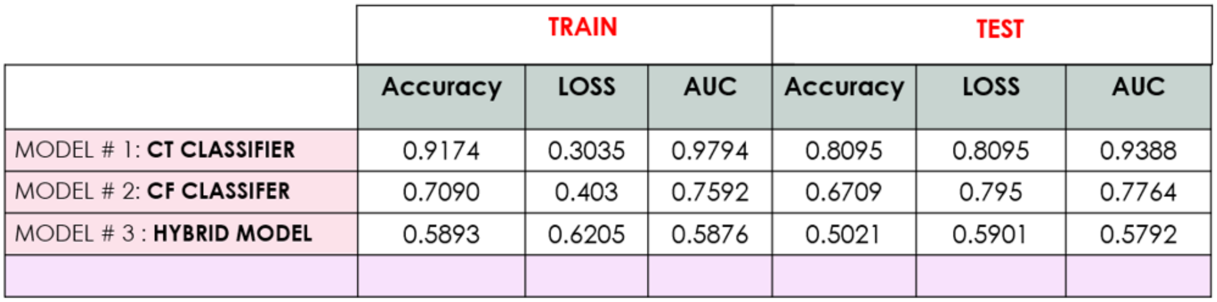


Figure 8 Evaluation Metric Results for Each Classifier Used within Research Project



	TRAIN			TEST		
	Accuracy	LOSS	AUC	Accuracy	LOSS	AUC
MODEL # 1: **CT CLASSIFIER**	0.9174	0.3035	0.9794	0.8095	0.8095	0.9388
MODEL # 2: **CF CLASSIFER**	0.7090	0.403	0.7592	0.6709	0.795	0.7764
MODEL # 3 : **HYBRID MODEL**	0.5893	0.6205	0.5876	0.5021	0.5901	0.5792

The figure above details the results obtained from the training and validation process for each prognosis predictor within this paper. The results clearly show that the model trained on CT images solely outperformed the other classifier models by attaining accuracy of 0.9174 and an AUC of 0.9388. A pattern was realized among the results, where the number of training data decreased so did the performance of the models. This is true for the model trained on solely clinical structured data (127 features) and the merging of previous probabilities (2 features) for inputs into the logistic regression.

DEPLOYMENT: A WEB-BASED APP

Classifier models detailed in this project were deployed as a web-based application, affording healthcare professionals overseeing covid-19 patients on admission to accurately determine the patients' prognostic status.

Figure 9 Homepage for Model Web-based Application

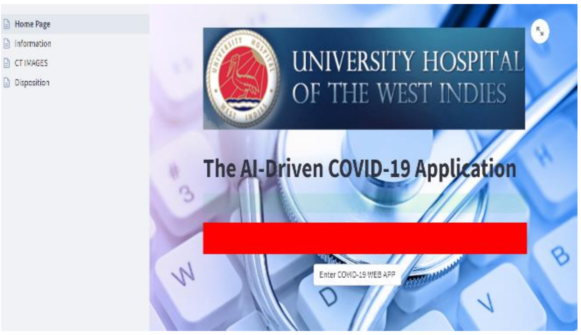

Home Page
Information
CT IMAGES
Disposition
UNIVERSITY HOSPITAL OF THE WEST INDIES
The AI-Driven COVID-19 Application
Enter COVID-19 WEB APP

Figure 10 Prognosis Prediction Using Structured Data

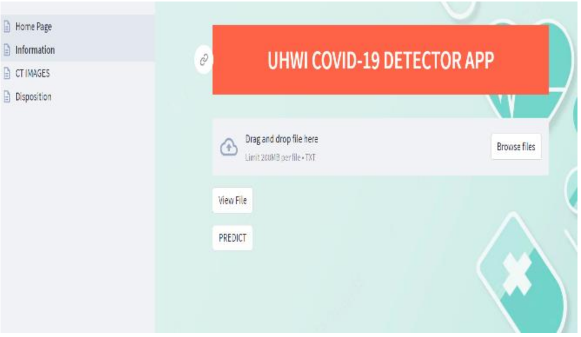


Figure 11 Prognosis Prediction Using CT Chest Images

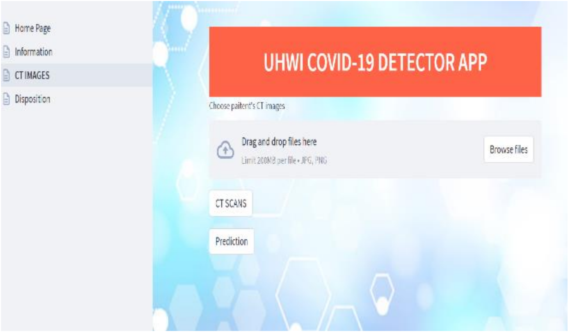
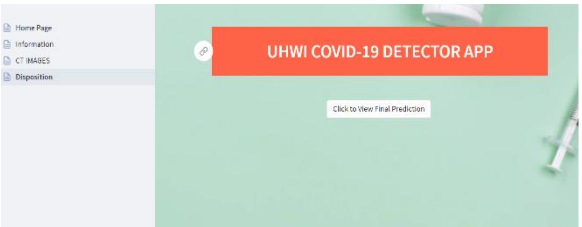


Figure 12 Final Prognosis Prediction, Probability Merger



CONCLUSION

In closing, the complexity and amount of biomedical data from patients admitted with covid-19 are typically hard to grasp, oftentimes leading to inaccurate prediction of a patient's prognosis; determining whether they will die or recover. Through this paper, we have been able to show that machine learning and deep learning models can be used to achieve this goal with varying levels of success. With that said, a successful prediction of a covid-19 patient's prognosis is heavily dependent of the amount of training data used. This paper has shown that a large dataset of images is able to successfully predict a patient's prognosis with accuracies of 91 percent. Having trained these models, they can be deployed as a web-based application for integration into clinical practice.

REFERENCES

Ning, W., Lei, S., Yang, J. *et al.* Open resource of clinical data from patients with pneumonia for the prediction of COVID-19 outcomes via deep learning. *Nat Biomed Eng* **4,** 1197–1207 (2020). https://doi.org/10.1038/s41551-020-00633-5

Roberts, M., Driggs, D., Thorpe, M. *et al.* Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nat Mach Intell* **3,** 199–217 (2021). https://doi.org/10.1038/s42256-021-00307-0

Wang, R., Jiao, Z., Yang, L. *et al.* Artificial intelligence for prediction of COVID-19 progression using CT imaging and clinical data. *Eur Radiol* **32,** 205–212 (2022). https://doi.org/10.1007/s00330-021-08049-8

Zhu J, Gallego B. Evolution of disease transmission during the COVID-19 pandemic: patterns and determinants. Scientific reports. 2021;11(1):1-9.