Integrated Model for Stratification of COVID-19 Patients

Martinus Boom
uc2024154464@student.uc.pt

Abstract

Using tabular and picture data, this project focuses on the stratification of COVID-19 patients, while a random forest (RF) model deals with numerical patient data and a convolutional neural network (CNN) scans ECG images. The approach aims to improve accuracy in predicting patient outcomes, and the need for hospitalization, by combining predictions from both models, and according to the results, the hybrid strategy outperformed models trained only on image or tabular data, achieving an accuracy of 84.17%.

1 Introduction

Precise patient classification is essential for using medical resources, especially in times of emergency like the COVID-19 pandemic, and determining whether to keep or dismiss patients who show symptoms is one big problem facing medical professionals. A simple study of clinical data is the base for this choice, which is made at the time of hospital admission. However, more advanced techniques are required to increase the accuracy of this stratification process due to the diversity of patient situations.

To deal with this problem, we develop an integrated model that makes use of both visual information in the form of ECG phase-space plots and numerical clinical data [8]. The model processes the numerical data using a Random Forest classifier and analyzes the binary ECG pictures using a Convolutional Neural Network (CNN), using the advantages of both methods, this strategy aims to improve patient outcome prediction. The integrated model aims to offer a more complete assessment by combining both data.

This study aims at constructing a hybrid machine-learning model which through clinical indicators and scanning of an ECG image can decide if a patient needs to be admitted or discharged. We aim to prove that there is an improvement in the performance of the combined model of CNNs and Random Forest when it's compared to the older forest models when it comes to performance and real-life use.

The report presents the main conclusions and how the integrated model was built, including the combination of CNNs with Random Forests for Patient data classification. Following that, the articles examine the architecture of the model and detail the results of the patient stratification model based on tabular and image data integration. In the last sections of the report, the characteristics of the model built are discussed, and the possibilities of its further development are indicated.

2 Related Work

Due to their high accuracy in identifying anomalies in ECGs, CNNs have developed significant use in medical imaging applications. At the same time, RF models have demonstrated their capacity to deal with structured tabular data, performing very well in classification tasks requiring test results.

Recent work on CNN architecture made for ECG image-based classification outperforms traditional techniques in sensitivity and accuracy in diagnosing cardiac issues as suggested in [1]. These findings emphasize the utility of utilizing CNNs for automatic diagnostic systems.

One paper applied an RF model to determine patient risk and forecast the outcome [2]. This decision has shown success in predicting disease growth in patients. Moreover, it reinforced their improvement in clinical decision-making.

Few studies have merged CNNs with other models for patient stratification, although CNNs are useful for classifying ECG images, sequential data is frequently handled by Long Short-Term Memory (LSTM) networks, which are made to recognize temporal patterns [3]. So there is potential to enhance stratification by utilizing both approaches, the combination of CNNs and Random Forests for tabular data categorization is yet not well studied.

There is a lack of research on combining CNNs for image analysis with RF models for tabular data, most methods utilize basic methods. To help with patient stratification, this study proposes an integrated model that processes patient data and ECG pictures together.

3 Materials and Methods

In this section, we describe the integrated model for patient stratification in terms of the datasets used, preprocessing methods applied, methodologies deployed, training of the model, and metrics to evaluate its performance.

3.1 Dataset and Preprocessing

To solve the patient stratification issue, this study used two linked datasets, COVID_numerics.csv and COVID_IMG.csv. The alignment of clinical and visual data for the same patient is ensured by the direct correspondence between each row of both datasets.

Six hundred samples containing eight clinical variables, such as gender, age, vaccination status, and vital signs, as well as a target variable that indicates the clinical decision 0 for returning home or 1 for hospitalization, are included in the COVID_numerics.csv dataset. 21x21 binary phase-space plots of ECG data are included in the COVID_IMG.csv dataset, which can provide spatial features.

Preprocessing Steps

To ensure consistency and increase efficiency, the following preprocessing steps were applied:

1. **Numerical Data Preprocessing:**
   - **Feature Selection and Cleaning:** Columns deemed extraneous, such as "Unnamed: 9" were deleted.
   - **Normalization:** Continuous variables and physiological data were transformed to achieve similar distributions and enhance model training performance.

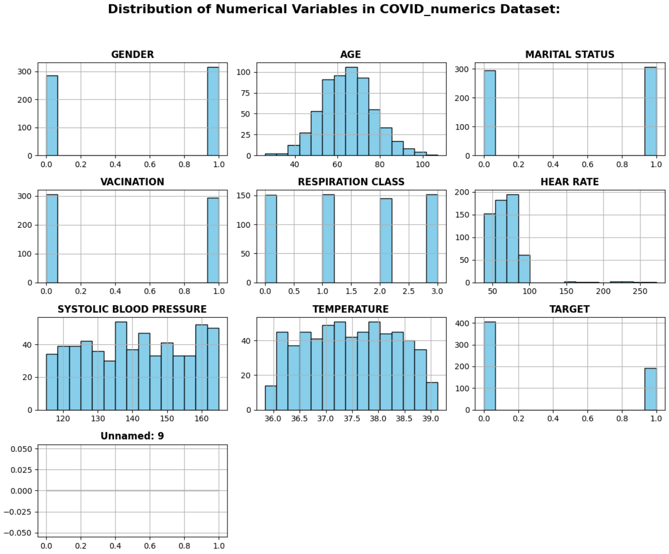

Figure 1: Distribution of Numeric Variables.

2. **Image Data Preprocessing:**
   - **Flattening:** ECG images of the initial structure of 21x21x1 pixels were transformed and sized into one-dimensional arrays of 441 bits for easier joining of the tabular data.
   - **Normalization:** To align the input distribution, the pixel values of the CNN images were adjusted to the ideal pixel values of 0 to 1 increasing the overall shift of the model.

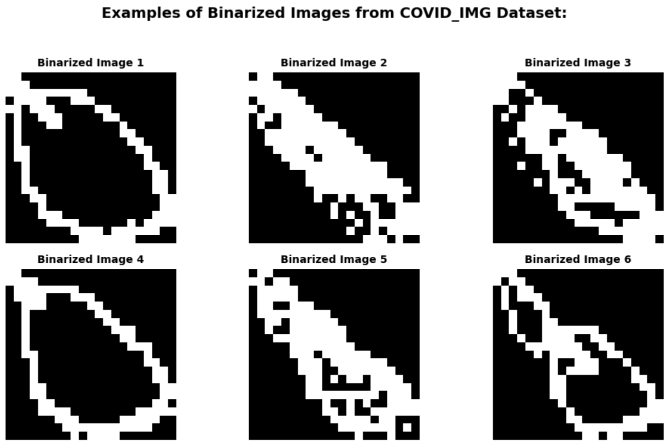

Figure 2: Examples of Binarized Images.

3. **Combined Data Preparation:**
   - **Concatenation:** The partially processed tabular data was merged with the ECG data to establish a consolidated dataset for balancing and passing into the model.
4. **Dataset Splitting:**
   - The data used for the training, validation, and testing of the model was divided into three sections in a ratio of 60%-20%-20%. The training set was used in the development of the integrated model. The validation set was simultaneously used to fine tuned parameters to limit overfitting. The assessment of the model and its generalizability was transcended through the use of the test set which was not only unbiased but also comprehensive.

3.2 Data Balancing

According to an evaluation of the preliminary datasets, a class distribution deficit was observed. It appears that there were significantly more non-critical patients than there were critical ones. Because of this skew, a model may be prone to predict mostly the non-critical patients while having a lack in predicting the current class. It can be observed in Figure 3 the class distributions for COVID_Numerics and COVID_IMG.

SMOTE

To balance this imbalance , the Synthetic Minority Oversampling Technique (SMOTE) was utilized, creating new samples for the minority class by combining existing samples. This maintains variance and mitigates the risk of overfitting to duplicate samples.

Application in this Study

In this work, SMOTE has only been applied to the training subset of the combined set which contained both the images and tabular data. This ensured that the synthetic examples created by SMOTE contained information about both new modalities and at the same time preserved the link between the image and tabular features. Since SMOTE was applied to the training dataset only, we did not change the inherent distribution of the classes in the validation and test datasets, thus allowing them to remain genuine evaluation datasets.

The stratified sampling was used to segregate the combined dataset into training (60%), validation (20%), and test (20%) sets with a view of preserving the proportions of the different classes that were in the validation and test sets.

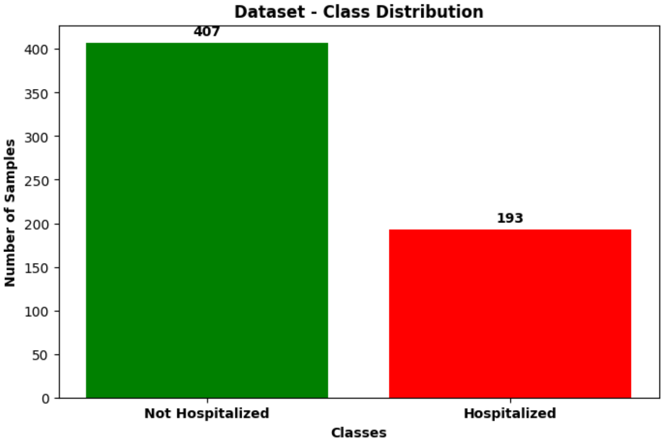

Figure 3: Class distribution in the Dataset.

3.3 Integrated Model Architecture

The structure of the model proposed in this project has two building blocks: one for working with image input and the other one for working with structured data. These building blocks work in unison to collect the best features from both modalities and later combine them for classification purposes. This two-branch structure guarantees that all the benefits of all the employed data types are used, thereby enhancing the accuracy of the prediction model.

Image Data Processing Branch

The image data processing sub-division employs a Convolutional Neural Network (CNN) which was tailored to image data set of 21x21 grayscale binary image. The structure of this branch is given in detail in Table 1. This architecture includes recurrent layers, pooling, and fully connected layers intended to learn the spatial components while processing ‘feature-abstraction’ to prepare for integration with the fusion layer.

This branch is specifically tailored for the image related tasks including the extraction of important features that help in classifying patients visually through binary means.

Numeric Data Processing Branch

The focus of the Numeric Data Processing unit is on the examination of structured quantitative clinical data like physiological and demographic characteristics This unit is described in Table 2. As a part of its processing, this unit utilizes a single scalar variable that is based on the prediction of a Random Forest trained using normalized numerical data. These predictions are later on turned into a feature vector with image based features in the integrated model.

Layer Type	Configuration	Purpose
Input	Shape: 21x21x1	Receive 21x21 single-channel grayscale images as input.
Conv2D	32 filters, 3x3, ReLU activation	Extract low-level spatial features from the image.
MaxPooling2D	Pool size: 2x2	Downsample feature maps, reducing computational complexity and preserving important features.
Conv2D	64 filters, 3x3, ReLU activation	Extract deeper spatial features.
MaxPooling2D	Pool size: 2x2	Further downsample feature maps.
Flatten	-	Convert 2D feature maps into a 1D feature vector for input into fully connected layers.

Table 1: CNN architecture for image data processing branch.

Layer Type	Configuration	Purpose
Input	Shape: 1x1	Accepts a single normalized numerical variable as input.
Dense	64 units, ReLU	Transforms the input into a high-dimensional feature vector.

Table 2: CNN architecture for numeric data processing branch.

Fusion Layer

The fusion layer architecture is illustrated in Table 3. This combined information is then used as the input for the final layer, which integrates the complementary information arising from both branches, guaranteeing that the ultimate model harnesses the advantages of both data modalities (for precise predictions).

This integrated architecture combines the features of the CNN image branch and the tabular data branch in the fusion layer which improves the accuracy of classification and interpretability of the model when compared to single input models.

Layer Type	Configuration	Purpose
Concatenation	Outputs of image and numeric branches	Combine features from both branches into a unified vector.
Dense	64 neurons, ReLU activation	Refine the combined feature set.
Dropout	Rate: 0.2	Prevent overfitting during training.
Output	1 neuron, Sigmoid activation	Predict the likelihood of hospitalization (binary classification).

Table 3: Feature fusion layer for combining image and numeric data.

3.4 Integrated Model Training

The merged model training procedure combines CNN branch capabilities with image data and RF forecasting from numeric data. Thus, this framework utilizes the harmonized characteristics of both modalities for improved predictions to be made.

RF Training and Feature Selection

The task of training was initiated with the tabular data in which a Random Forest (RF) model was used. Predictions made from this model were provided as input to the integrated model afterward.

- **Feature Reduction:** In order to refine the RF model, seven of the features determined to be the most valuable were safeguarded while the rest were discarded. This determination stemmed from the analysis that the inclusion of the seven features had resulted in a reduction of log-loss. This step ensured the model concentrated on the best predictors rather than being overly complicated.

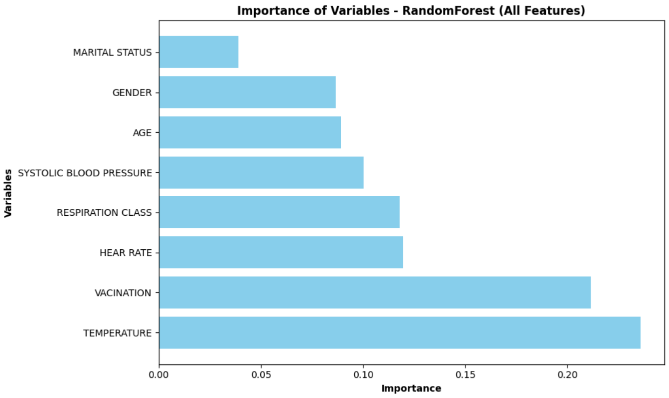

Figure 4: Importance of All Variables.

Integrated Model Training Framework

The integrated model was trained on concatenating CNN features extracted from image data and RF predictions based on tabular data. These were combined in the fusion layer and then trained jointly to get the final prediction.

- **Training Function:** The training function also focused on minimizing the accuracy-binary cross-entropy loss as a secondary evaluation metric. We linearized both image features and RF predictions and combined them to ensure data modality and seamless integration.
- **Callbacks:**
  - **Early Stopping:** Watched the checking loss to stop training when no gain was seen for 10 straight rounds, cutting down on useless work and stopping overfitting.
  - **Learning Rate Scheduler:** Changed the learning speed up little by little by cutting it down by half when check loss stayed flat for 5 rounds, making it smoother to meet and stopping strong swings in learning.

3.5 Evaluation Metrics

Accuracy and loss were selected as the main criteria to assess the active learning framework performance. These metrics were produced at each iteration ensuring a consistent evaluation, among others:

1. **Accuracy**
2. **Loss**
3. **Precision**
4. **Recall**
5. **F1-Score**
6. **ROC AUC**

4 Results and Discussion

4.1 Results

Model	Accuracy	Log Loss
Random Forest (All Features)	0.8667	0.3262
Random Forest (Top 7 Features)	0.8500	0.3188
Integrated Model (CNN + RF)	0.8417	5.7069

Table 4: Performance Metrics for Random Forest and Integrated Model

Class	Precision	Recall	F1-Score
Not Hospitalized	0.87	0.90	0.89
Hospitalized	0.77	0.71	0.74

Table 5: Precision, Recall, and F1-Score for the Integrated Model

Performance Across Iterations

Figures 5 and 6 represents the training and validation accuracy and loss patterns for the integrated model framework across iterations.

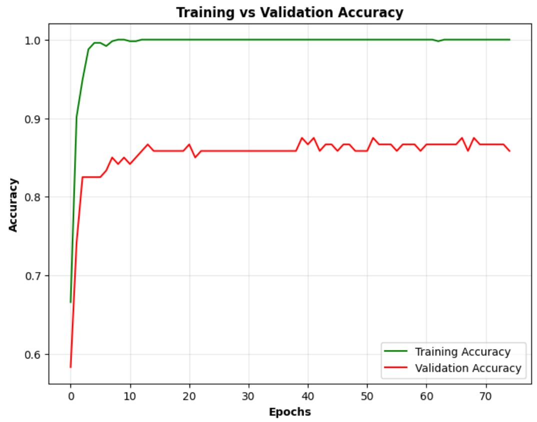

Figure 5: Training and validation accuracy.

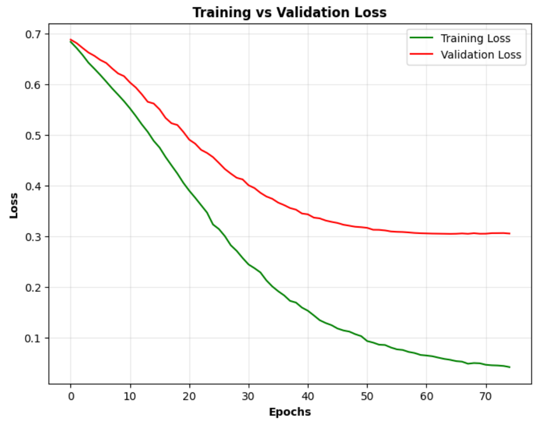

Figure 6: Training and validation loss.

Confusion Matrix

Figure 7 show the confusion matrix for the integrated model.

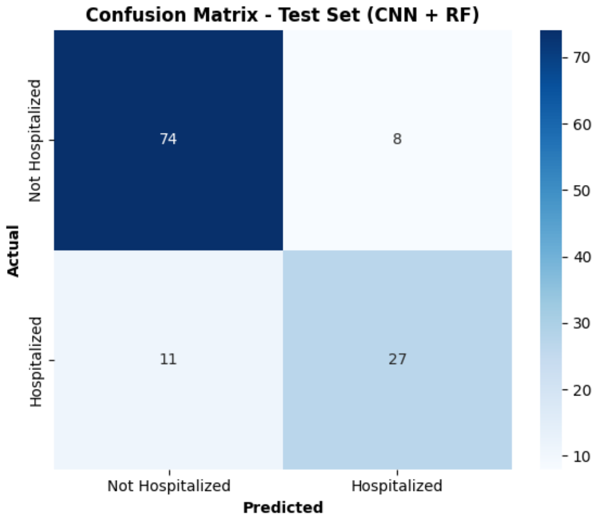

Figure 7: Confusion Matrix of Integrated Model.

ROC AUC

Figure 8 shows the sensitivity and specificity trade-off of the integrated model with accuracy value of 0.8986.

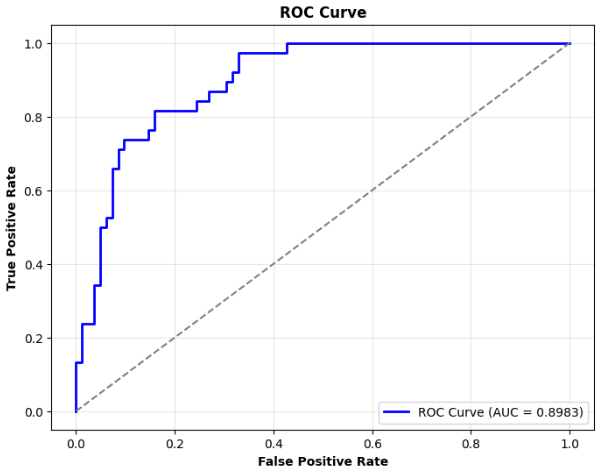

Figure 8: ROC Curve of Integrated Model.

4.2 Discussion

The results in Section 4.1 correspond to the classic approach by integrating CNN for image data, and for numerical data, the random forest predictions.

Due to the complexities within the dataset, the random forest model trained on the complete feature set was able to achieve high accuracy (86.67%) and relatively low log loss (0.3262), indicating its potential as a powerful tool in the latter part of our discussion where tabular clinical data will be of prime importance. When we limited the features to the most significant seven we saw a small reduction in accuracy (85.00%) and a slight improvement in log loss (0.3188). This suggests feature reduction does a good job of filtering out noise while retaining predictive power.

The selection of these seven features slightly improved log loss following feature reduction indicates that the model's probabilistic predictions turned out to be more confident and reliable. So, from the chosen features, the key signs were Temperature and Vaccination. The strong value placed on these features to the idea that temperature is seen as a big clinical sign because fever is often one of the main signs of COVID-19. The next most helpful feature was Vaccination, which shows how vaccines help in preventing serious issues and lowering hospital stays. These results show how the model took in important health factors and also showed why it's key to include them in guesses for COVID-19 hospital visits.

With an overall accuracy of 84.17%, it still showed the potential of having its images and numeric data combined thanks to the extreme reliance of random forests on structured numerical information and CNNs on spatial features. Nonetheless, this dual-modality approach had its downsides. The random forest model which was trained on only tabular data managed to outshine the integrated model scoring

85% in accuracy. Moreover, concerning log loss, the hybrid model's log loss of 5.7069 is larger in comparison to the reduced random forest model which has a log loss of 0.3188 these numbers further enhance the prediction of the hybrid model. All in all, these challenges suggest that the use of image data along with data that is tabular can complicate an otherwise simple approach to making predictions.

The integrated model has some advantages and constraints which are more visible through the application of precision, recall, and F1 measures. For the majority class (Not Hospitalized), high precision (0.87) and recall (0.90) enable the model to predict a non-hospitalized case with low false positive rates. On the other hand, in the case of the minority class (Hospitalized), the scores for precision (0.77) and recall (0.71) are relatively poor since the model faces obstacles in designing algorithms that can predict the number of cases that are Hospitalized and has more of false negative. The "F1-score" metric is more balanced as it captures both precision and recall balance it, and was found to be equal to 0.89 for the majority class and 0.74 for the minority case. These classes have been far apart to suggest a "good" performance for the class with more instances but a "bad" performance for the class with fewer instances. The effectiveness of the model's overall predictive ability would have been better if the class imbalance, which is also apparent, and the feature representation for the minority class were at par with other factors.

With 74 true positives and 8 false positives, the confusion matrix demonstrates how well the integrated model identified the majority class (not hospitalized), otherwise, it had trouble with the hospitalized minority class, though, since it produced 11 false negatives and 27 true positives. Even after using SMOTE, this discrepancy most likely results from a class imbalance in the original dataset, which could have caused biased feature learning, and, merging CNN and random forest predictions might have added noise or made it harder for the model to correctly identify the minority class.

Last but not least, while the ROC-AUC score of 0.8986 suggests strong discrimination ability, it fails to deal with the serious practical consequences of false negatives in the hospitalized class, which may result in cases being ignored that need urgent attention.

5 Conclusion

To summarize, the integrated model in question was shown to be feasible for the integration of multi-modal data for medical applications, utilizing CNN for image based spatial features and random forests for numeric data. However, its performance showed considerable weaknesses. Though the model reached 84.17% overall accuracy, the comparatively greater log loss (5.7069) concerning the random forest model (0.3188) indicated the issues in probabilistic calibration and confidence in the predictions of the model. This constraint also indicates the degree of difficulty brought about by the combination of different data types.

Not being hospitalized and being hospitalized is still a classification challenge, and the imbalance between precision and recalls for both earlier stems from the surgery, given there is a class imbalance. Not getting hospitalized was correctly diagnosed by the model, however, many patients who were supposed to be in the hospitalized category were instead diagnosed as false negatives. There is a need to address this area because, due to improper classification, crucial oversights in real life may occur, stressing the need for model configurations to have a more favorable result for a patient.

Further refinements in image-numeric fusion, improving model calibration, and mitigating class imbalance through dynamic re-weighting or sophisticated data augmentation will go a long way to improving this work. This would very significantly enhance the reliability of the model by much more clinically actionable predictions.

References

[1] IEEE Xplore. ECG Image Classification Using CNN for Cardiac Disease Diagnosis. Available at: https://ieeexplore.ieee.org/document/10796798

[2] BMC Medical Research Methodology. Dynamic Clinical Risk Prediction Using Random Forest for Multivariate Outcomes. Available at: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-019-0863-0

[3] International Information and Engineering Technology Association (IIETA). ECG Image Classification Integrated with Random Forest Models for Patient Stratification. Available at: https://iieta.org/journals/isi/paper/10.18280/isi.290532

[4] Kaggle Discussions. SMOTE: Imbalanced Data. Available at: https://www.kaggle.com/discussions/general/220862

[5] Kaggle Discussions. Accuracy vs Loss Conflict. Available at: https://www.kaggle.com/discussions/general/220823

[6] Kaggle Code. Random Forest Classifier Tutorial. Available at: https://www.kaggle.com/code/prashant111/random-forest-classifier-tutorial

[7] V7 Labs. Multimodal Deep Learning Guide. Available at: https://www.v7labs.com/blog/multimodal-deep-learning-guide

[8] Martinus Boom. GitHub Repository: ML COVID Patient Stratification. Available at: https://github.com/MartinusBoomUc2024154464/ML_CovidPatientStratification.git