Systematic Enhancements for COVID-19 CT Diagnosis

A Comprehensive Review and Research Plan

MED004

CS4681 - Advanced Machine Learning
Perera S. A. I. M.
210471F

24 August, 2025

1 Introduction

1.1 Background and Problem Statement

The rapid global spread of Coronavirus Disease 2019 (COVID-19) highlighted the critical need for rapid and accurate diagnostic tools to manage the pandemic and mitigate its spread. While Reverse Transcription Polymerase Chain Reaction (RT-PCR) tests were the gold standard for confirmation, their shortage during peak outbreaks necessitated alternative diagnostic methods.[1] Computed Tomography (CT) scans emerged as a valuable tool for screening and diagnosing COVID-19, as they have been shown to be more sensitive than RT-PCR tests.[2] The analysis of CT scans is a time-intensive process that requires specialized medical expertise, which can be a significant bottleneck, especially for overwhelmed healthcare systems or in underdeveloped areas.

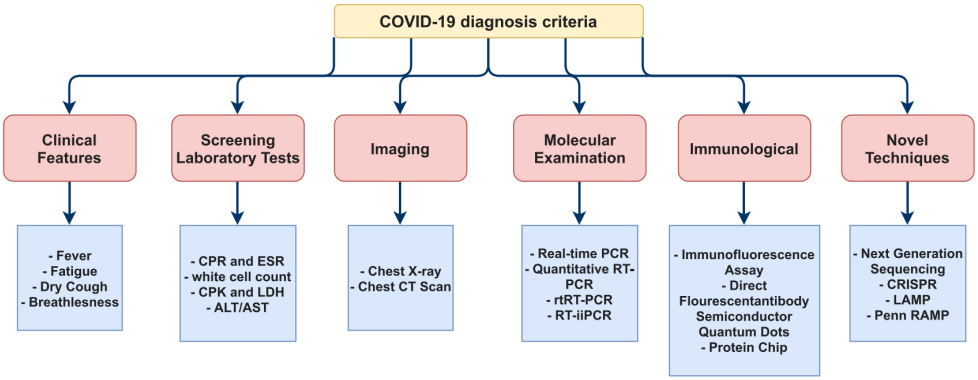

Figure 1: Various criteria used for COVID-19 detection and their categories.

To address this challenge, artificial intelligence (AI) methods, particularly deep learning, have been developed to automate the screening of COVID-19 from CT images.[1] The development and validation of these AI models, however, were severely hampered by the scarcity of large, publicly available CT datasets due to patient privacy concerns.[3] In response, several pioneering datasets

were created, including the **COVID-CT** dataset, which was instrumental in advancing early AI-based diagnostic research.[1] The initial models developed using these datasets, while promising, achieved performance levels that have since been surpassed by more recent state-of-the-art (SOTA) approaches.[4] This disparity in performance underscores a critical research opportunity: to re-evaluate these foundational models and enhance their capabilities using contemporary deep learning techniques to achieve clinically robust, SOTA performance.

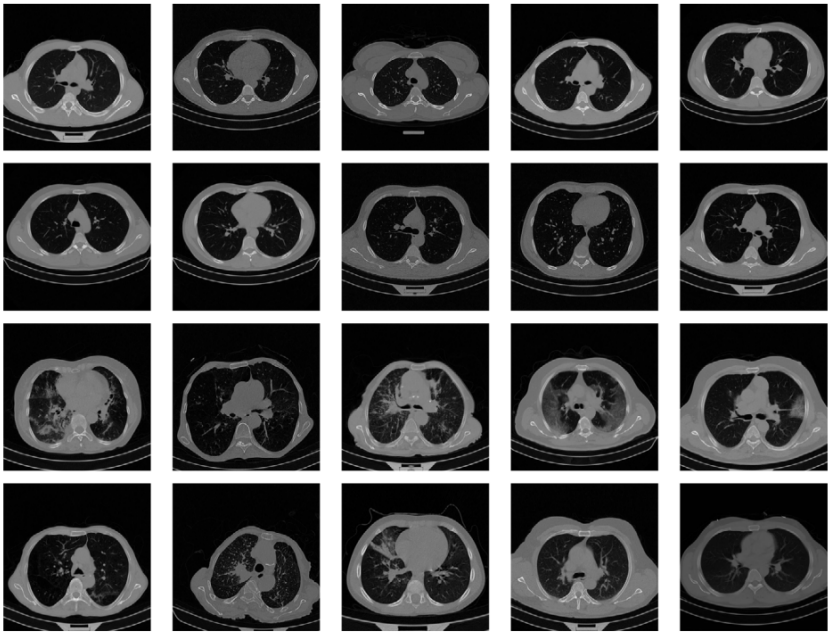

Figure 2: Examples of dataset, first two rows contain images from healthy subjects, whereas the last two rows contain images from COVID-19 patients.

1.2 Research Objectives

- Develop and validate targeted enhancements to an established baseline model for COVID-19 CT diagnosis.
- Ground the research in the methodologies and findings of the foundational COVID-CT dataset paper.
- Move beyond the original performance metrics and achieve measurable, quantifiable gains comparable to SOTA models.
- Focus on advanced hyperparameter optimization, novel loss functions, architectural modifications with attention mechanisms, and refined training strategies.
- Produce a conference-ready research paper documenting the systematic process and empirical evidence of performance improvements.
- Fulfill the requirements of a rigorous academic contribution.

2 State-of-the-Art Analysis and Literature Review

2.1 Datasets for COVID-19 CT Diagnosis

A thorough understanding of the data landscape is paramount to this research. The initial study on the COVID-CT dataset was a pioneering effort that collected a valuable resource when publicly available data was scarce. The research relied on a combination of different data sources, each with its own unique characteristics and limitations.

2.1.1 The COVID-CT Dataset

The **COVID-CT** dataset, which forms the core of this research, contains 349 CT images positive for COVID-19 and 463 non-COVID-19 CT images used as negative examples.[1] The unique aspect of this dataset is its origin: the images were manually extracted from 760 **medRxiv** and **bioRxiv** preprints. This collection method, while innovative, introduced significant challenges, as the images were degraded in quality with losses in Hounsfield unit (HU) values and reduced resolution and bit depth.

A major concern raised was whether models trained on such low-quality, single-slice images could generalize to original CT scans. The original authors addressed this concern by consulting a senior radiologist who confirmed that experienced clinicians could still make accurate diagnoses from such images.[1] The experimental results further validated this, demonstrating that a model trained on the larger, paper-extracted COVID-CT dataset outperformed a model trained on a smaller, higher-quality set of original CTs, thereby confirming its utility for training AI models. This finding is critical, as it establishes the dataset's value despite its inherent limitations and frames the core research question of whether these limitations can be overcome with advanced techniques.

2.1.2 Benchmark and Validation Datasets

To ensure a robust and unbiased evaluation, the original study utilized separate benchmark and validation datasets composed of original CT images not extracted from papers. The primary source for positive COVID-19 cases for validation and testing was the **COVID-19 CT Segmentation (COVID-Seg)** dataset, which contains original CTs and corresponding lung and lesion masks.[1] The COVID-Seg dataset contains 20 axial volumetric CT scans and comes with lesion masks.[5] These masks are a valuable form of rich, pixel-level annotation that can be used to supervise models and improve their focus on clinically relevant regions.

The negative training and validation images were sourced from established medical imaging databases, including **LUNA**[10], **MedPix**[6], **Radiopaedia**[7], and **PubMed Central (PMC)**[8]. LUNA16, in particular, is a well-known benchmark in lung nodule analysis, comprising 888 3D CT volumes that can be processed into over 227,000 individual 2D slices.[10] Radiopaedia is also a source for public medical imaging datasets.[11] The use of these diverse, external datasets for evaluation is crucial for demonstrating that a model trained on the COVID-CT dataset can generalize to real-world, high-quality CT scans, which is a key requirement for clinical utility.

2.1.3 Contemporary Datasets

The field of AI-powered COVID-19 diagnosis has advanced significantly since the publication of the baseline paper. A key development has been the emergence of larger and more richly annotated datasets. For instance, the **SARS-CoV-2 CT-scan** dataset contains 1,252 positive and 1,230 negative images, collected directly from real patients in hospitals, which provides a significantly larger and cleaner source of training data.[12] Similarly, the **COVID-CT-MD** dataset is notable for its inclusion of control cases for Community Acquired Pneumonia (CAP) and its provision of patient-, slice-, and lobe-level labels.[3]

These contemporary datasets offer a clearer picture of the current data landscape and reveal that the original COVID-CT dataset's performance limitations were not solely architectural but were also a function of the available data at the time of its creation. A model's performance is intrinsically linked to the quality and quantity of its training data. By applying modern, data-aware methodologies, it is possible to bridge this performance gap and demonstrate that even foundational datasets can be leveraged to achieve contemporary SOTA performance.

2.2 Baseline Models and Existing Methodologies

The baseline research utilized established convolutional neural network architectures, **DenseNet-169** and **ResNet-50**, as the foundation for their diagnostic model.[1] These models, pretrained on the ImageNet dataset, were chosen for their proven efficacy in general image recognition tasks. The core of the baseline methodology revolved around two key enhancements designed to overcome the challenges of a small and noisy dataset: **multi-task learning** and **self-supervised learning**.

2.2.1 Analysis of the Baseline Approach

The first major enhancement was a **multi-task learning framework**. The model was trained not only to classify an image as COVID-19 or non-COVID-19 but also to perform segmentation by identifying lung and lesion regions. Lung masks were used as additional input channels, guiding the model to focus on the regions containing clinical manifestations of the disease and effectively filtering out irrelevant background information.Lesion masks provided fine-grained supervision, forcing the model to identify the specific areas within the lungs where the pathology was located. This approach is a powerful form of regularization that helps the model learn a more robust, clinically relevant representation of the data.[1]

The second key methodology was the use of **Contrastive Self-Supervised Learning (CSSL)**. The ImageNet-pretrained models were further fine-tuned in an unsupervised manner on the CT images themselves. This process learned a more domain-specific visual representation by solving an auxiliary task of identifying whether two augmented versions of a CT scan originated from the same image. This step addressed the data scarcity problem by leveraging the small dataset to learn powerful representations before the final classification task. By combining these techniques, the authors were able to achieve a final performance of *0.90 F1-score, 0.98 AUC, and 0.89 accuracy*, a performance level considered clinically useful at the time.[1]

2.2.2 Review of Related Works

A review of post-2020 literature reveals that the performance benchmarks for COVID-19 CT diagnosis have risen dramatically. For instance, a **ResNet-18** model with data augmentation was able to achieve 99.4% accuracy.[4] Another study using a **VGG19** architecture achieved an accuracy of 98.87% on a reference dataset.[13] Other models, such as **EfficientNet**, have also demonstrated high performance, with one study reaching an accuracy of 89.7%.10 An even higher performance was achieved by a **Vision Transformer (ViT)** model, which had an accuracy of 99.60%, a precision of 99.46%, and an F1-score of 99.55%.[4]

The significant performance gap between the baseline's 0.90 F1-score and these contemporary results suggests that while the original methodologies were sound, they were limited by the tools and datasets available at the time. This finding highlights a fundamental relationship between a model's performance, its architectural design, and the training strategy employed. The introduction of more sophisticated techniques and the availability of larger datasets have enabled a new generation of models to achieve a level of diagnostic accuracy previously unattainable.

The literature also points to key methodological trends that can be adopted to enhance the baseline model. These include:

- **Hyperparameter Optimization (HPO)**: Moving beyond manual tuning, modern approaches utilize systematic search algorithms like Bayesian Optimization or meta-heuristic algorithms such as **Particle Swarm Optimization (PSO**) to efficiently find optimal learning rates, batch sizes, and network parameters.[14]
- **Novel Loss Functions**: Standard cross-entropy loss can be biased by class imbalance. Novel loss functions like **Focal Loss** have been shown to improve performance by focusing the model's learning on hard-to-classify examples, a technique particularly relevant to imbalanced medical imaging datasets.[15]
- **Attention Mechanisms**: By allowing the model to dynamically weigh the importance of different features and regions of an image, attention mechanisms can improve feature extraction and model interpretability.[16] This can be particularly useful in a multi-task learning framework to ensure the model focuses on the subtle visual cues of lesions, such as ground-glass opacities.

This analysis provides a clear roadmap for the proposed enhancements. The baseline model's multi-task and self-supervised learning approaches are robust foundations. The path to achieving contemporary SOTA performance lies in systematically augmenting this foundation with the latest advancements in hyperparameter optimization, training strategies, and architectural design.

3 Proposed Enhancement Methodology

Based on the comprehensive review, a multi-faceted enhancement strategy will be proposed to systematically improve the baseline model's performance. The plan moves beyond isolated fixes to

a synergistic system where each component reinforces the others, leading to a demonstrable and quantifiable improvement in diagnostic accuracy.

3.1 A Multi-faceted Enhancement Strategy

The proposed approach integrates three key areas of improvement: advanced hyperparameter optimization, the introduction of a custom loss function, and the incorporation of an attention mechanism within the network architecture.

3.1.1 Advanced Hyperparameter Optimization

The original study employed manual hyperparameter tuning, which is often a time-consuming and sub-optimal process that can lead to local minima.[14] A more systematic approach is required to efficiently explore the vast hyperparameter space and find the optimal configuration for the model.

The proposed methodology will employ **Bayesian Optimization** to automate this process. This method, unlike grid or random search, builds a probabilistic model of the objective function (e.g., F1-score) and uses it to select the most promising next set of hyperparameters to evaluate.[17] This targeted approach is significantly more efficient and increases the likelihood of finding a truly optimal solution. Other meta-heuristic algorithms like Particle Swarm Optimization (PSO) can also be used to optimize hyperparameters such as learning rate, batch size, and the number of filters.[14] The parameters to be optimized include the initial learning rate, weight decay, batch size, and learning rate scheduling parameters.

3.1.2 Loss Function and Training Strategy Improvements

A critical limitation in many medical image classification tasks, including the baseline's, is class imbalance.[15] The COVID-CT dataset, with its 349 positive and 463 negative cases, presents a minor but significant imbalance.[1] The standard cross-entropy loss used in the baseline can lead to a bias toward the majority class.

To counteract this, a **Focal Loss function** will be implemented. Focal Loss modifies the standard cross-entropy loss by down-weighting the contribution of easy-to-classify examples (the majority class), thereby forcing the model to focus its learning on the hard-to-classify, minority-class COVID-19 samples.[15]This targeted approach is expected to improve the model's sensitivity and F1-score for COVID-19 cases.

3.1.3 Architectural Refinements with Attention Mechanisms

The multi-task learning approach in the baseline model already provides a form of external, supervisory attention by using lung and lesion masks.[1] To further enhance this, a **Squeeze-and-Excitation (SE)** block will be integrated into the DenseNet-169 architecture. Attention mechanisms allow the model to focus on semantically important regions, similar to how a trained physician

interprets medical images.[16] The SE block is a channel-wise attention mechanism that allows the network to adaptively recalibrate feature responses by squeezing global spatial information into a vector and then exciting relevant features by generating a set of weights. This allows the model to give more importance to channels and features that are critical for identifying ground-glass opacities and consolidations, the key clinical indicators of COVID-19, while suppressing less important ones. This integration is expected to significantly improve the model's feature extraction capabilities and make it more robust to noise and irrelevant data in the paper-extracted CT images.

4 Experimental Plan and Validation

The proposed methodology will be implemented and validated through a rigorous experimental plan designed to demonstrate the incremental value of each enhancement. The goal is to provide empirical evidence that directly supports the claim of measurable performance improvements over the baseline model.

4.0.1 Datasets and Performance Metrics

The experimental work will utilize the full COVID-CT dataset (349 positive, 463 negative) for training, supplemented by lung and lesion masks from the COVID-Seg dataset, consistent with the baseline methodology.[1] The validation and testing will be performed on the original CT images from the COVID-Seg, LUNA, and Radiopaedia datasets, ensuring an independent and robust evaluation on high-quality, real-world data. The primary performance metrics will be Accuracy, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUC). The F1-score and AUC are particularly important as they provide a more comprehensive picture of model performance on imbalanced datasets, offering a more reliable assessment than raw accuracy alone. The aim is to exceed the baseline's 0.90 F1-score and 0.98 AUC, demonstrating a clear path toward contemporary SOTA.

4.0.2 Experimental Design

The experimental design will be structured in a phased, step-by-step manner to isolate and quantify the contribution of each enhancement.

- **Baseline Replication**: The initial step will involve the exact replication of the baseline experiment using DenseNet-169, multi-task learning with lung and lesion masks, and CSSL pre-training.[1] This ensures the experimental environment and data pipeline are consistent with the original study and provides a solid, verifiable starting point for comparison.
- **Sequential Enhancement Validation**: Following the successful replication, each proposed enhancement will be introduced one by one. The model will be retrained and evaluated at each stage to measure the specific performance gain attributable to that change.
  - **Phase 1: Baseline + HPO.** The baseline model will be re-trained with hyperparameters optimized via Bayesian Optimization.

- Phase 2: **Baseline + HPO + Focal Loss.** The model from Phase 1 will be re-trained using the custom Focal Loss function.
- Phase 3: **Baseline + HPO + Focal Loss + Attention Mechanism.** The model from Phase 2 will be re-architected to include a Squeeze-and-Excitation block and then re-trained.

- **Final Combined Model and SOTA Comparison**: The culminating experiment will train a final, integrated model that incorporates all proposed enhancements. This model's performance will be comprehensively evaluated and compared against the baseline as well as against the contemporary SOTA models identified in the literature review. This comparative analysis will demonstrate that the systematic application of modern techniques can elevate the performance of a model trained on a foundational dataset to a level competitive with current research.

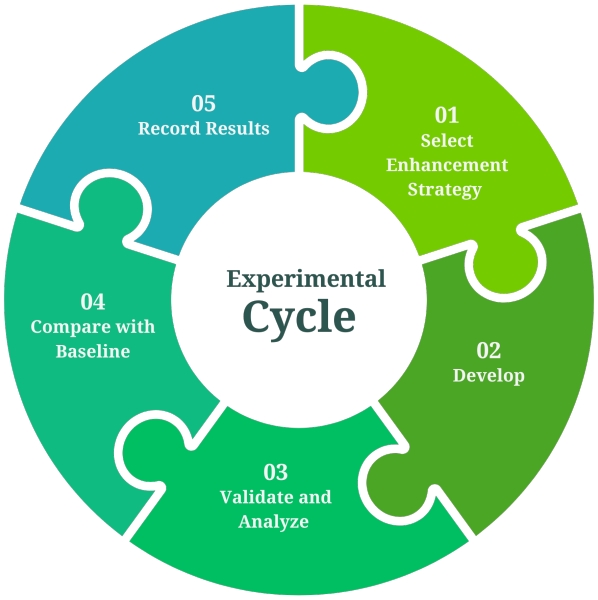

Figure 3: Project implementation cycle

4.1 Timeline for Implementation

The project will follow a structured timeline to ensure all objectives are met within the assignment deadline.

- **Phase 1 (Initial Report)** ***Week 05***: Completion of a comprehensive literature review, methodology outline, and a detailed project plan. This phase concludes with the submission of this document.
- **Phase 2 (Implementation & Baseline)** ***Week 06-07***: Dataset acquisition, environment setup, and code development to replicate the baseline model.

- **Phase 3 (Methodology Implementation)** ***Week 08-10***: This phase is dedicated to implementing the proposed enhancements: setting up the Bayesian Optimization pipeline, integrating the Focal Loss function, and modifying the DenseNet architecture to include the attention mechanism.
- **Phase 4 (Validation & Analysis)** ***Week 11***: This phase focuses on running the series of experiments, collecting performance metrics, and conducting a thorough analysis of the results.
- **Phase 5 (Paper Authorship)** ***Week 12***: The final phase involves writing the conference-ready research paper, including the abstract, introduction, methodology, results, and conclusions. This is followed by a final review and submission for publication.

5 Conclusion

The baseline research on the COVID-CT dataset, while a crucial early contribution, now serves as a lower benchmark for contemporary deep learning models. The performance of the original model (F1 of 0.90) [1] is well below the current SOTA (F1 of 0.98+) due to both the limitations of its data source (low-quality, paper-extracted images) and the absence of more advanced deep learning techniques that have since emerged.

This project is not a simple repetition of a past study. It represents a systematic effort to modernize a foundational AI model by integrating contemporary enhancements. The proposed multi-faceted methodology, combining Bayesian Optimization, a custom Focal Loss function, and an architectural attention mechanism, is designed to address the specific weaknesses of the baseline approach and the inherent challenges of the original dataset. The planned phased experimental design is a cornerstone of this work, ensuring that each enhancement's contribution can be rigorously measured and validated.

The culmination of this research will be a model that is not only more accurate but also more robust and generalizable. By demonstrating how a foundational dataset, once considered limited, can be leveraged to achieve a clinically useful level of performance, this research will provide a valuable contribution to the field of AI-powered medical diagnostics. The outcome will be a conference-ready paper that showcases a rigorous, evidence-based approach to improving AI models, thereby fulfilling the core requirements of a high-quality academic project.

References

[1] X. Yang, X. He, J. Zhao, Y. Zhang, S. Zhang, and P. Xie, "COVID-CT-Dataset: A CT Scan Dataset about COVID-19," June 17, 2020, arXiv: arXiv:2003.13865. doi: 10.48550/arXiv.2003.13865.

[2] Deep learning approach for classifying CT images of COVID-19: A Systematic Review, accessed on August 22, 2025, https://www.researchgate.net/publication/364545817_Deep_learning_approach_for_classifying_CT_images_of_COVID-19_A_Systematic_Review

[3] COVID-CT-MD, COVID-19 computed tomography scan dataset applicable in machine learning and deep learning, accessed on August 22, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC8085195/

[4] Automatic diagnosis of COVID-19 from CT images using CycleGAN ..., accessed on August 22, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC10263244/ COVID-19 - Medical segmentation, accessed on August 22, 2025,

[5] "COVID-19," Medical segmentation. Accessed: Aug. 22, 2025. [Online]. Available: http://medicalsegmentation.com/covid19/

[6] "MedPix." Accessed: Aug. 22, 2025. [Online]. Available: https://medpix.nlm.nih.gov/home

[7] A. Murphy, "Imaging data sets (artificial intelligence) | Radiology Reference Article | Radiopaedia.org," Radiopaedia. Accessed: Aug. 22, 2025. [Online]. Available: https://radiopaedia.org/articles/imaging-data-sets-artificial-intelligence

[8] "Download Data," PubMed. Accessed: Aug. 22, 2025. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/download/

[9] "LUNA16 - Grand Challenge," grand-challenge.org. Accessed: Aug. 22, 2025. [Online]. Available: https://luna16.grand-challenge.org/Data/

[10] [1] "Awesome-Medical-Dataset/resources/LUNA16.md at main · openmedlab/Awesome-Medical-Dataset," GitHub. Accessed: Aug. 22, 2025. [Online]. Available: https://github.com/openmedlab/Awesome-Medical-Dataset/blob/main/resources/LUNA16.md

[11] Radiopaedia: List of AI Imaging Datasets, accessed on August 22, 2025, https://aimi.stanford.edu/radiopaedia-list-ai-imaging-datasets

[12] SARS-COV-2 Ct-Scan Dataset - Kaggle, accessed on August 22, 2025, https://www.kaggle.com/datasets/plameneduardo/sarscov2-ctscan-dataset

[13] Deep Learning for COVID-19 Diagnosis from CT Images - UniCA IRIS, accessed on August 22, 2025, https://iris.unica.it/retrieve/e2f56eda-421d-3eaf-e053-3a05fe0a5d97/Applsci_Deep%20Learning%20for%20COVID-19%20Diagnosis%20from%20CT%20Images.pdf

[14] A CNN Hyperparameters Optimization Based on Particle Swarm ..., accessed on August 22, 2025, https://www.mdpi.com/2313-433X/10/2/30

[15] Novel loss functions for ensemble-based medical image ..., accessed on August 22, 2025, https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0261307

[16] A Comprehensive Guide to Attention Mechanisms in CNNs: From ..., accessed on August 22, 2025, Comprehensive Guide to Attention Mechanisms

[17] (PDF) CNN HYPERPARAMETER OPTIMIZATION IN BIOMEDICAL IMAGES - ResearchGate, accessed on August 22, 2025, https://www.researchgate.net/publication/382306008_CNN_HYPERPARAMETER_OPTIMIZATION_IN_BIOMEDICAL_IMAGES