Journal of Artificial Intelligence and Metaheuristic

♦ ISSN: 2833-5597 Vol: 12, No.01, 2027 ♦ PP. 18–41

♦ Article DOI: https://doi.org/10.54216/JAIM.120103

FULL LENGTH ARTICLE

Feature Selection and Hyperparameter Optimization of Convolutional Neural Networks Using the Ocotillo Optimization Algorithm for Pneumonia Detection from Chest X-Ray Images

Mona Yassen [1,*] **Ancy Cheriyan** [2]

[1] Computer Science and Intelligent Systems Research Center, Blacksburg 24060, Virginia, USA
[2] School of Artificial Intelligence, Bahrain Polytechnic, PO Box 33349, Isa Town, Bahrain

Emails: mona@jcsis.org ; ancy.cheriyan@polytechnic.bh

Received: April 19, 2026 Revised: June 28, 2026 Accepted: August 12, 2026 ⋆ Corresponding author

ABSTRACT

Pneumonia diagnosis from chest X-ray images remains clinically important because subtle radiographic abnormalities can be difficult to assess consistently under high workload or limited expert availability. This study proposes an optimization-guided diagnostic framework that integrates the Ocotillo Optimization Algorithm (OcOA) with a Convolutional Neural Network (CNN) for binary pneumonia classification. Rather than introducing OcOA or CNN as new algorithms, the contribution lies in using OcOA in two complementary stages: feature selection over CNN-derived deep representations and hyperparameter optimization of the CNN classifier. The proposed pipeline was evaluated on the Chest X-Ray Images (Pneumonia) dataset, which contains pediatric radiographs labeled as Normal or Pneumonia. In the feature-selection stage, binary Ocotillo Optimization Algorithm (bOcOA) achieved the lowest average error, with a value of 0.523417. After feature selection, the CNN baseline achieved the best classification performance among the evaluated machine-learning models, reaching an accuracy of 0.94985. In the hyperparameter optimization stage, OcOA+CNN achieved the strongest optimized performance, with an accuracy of 0.989939. These findings indicate that OcOA can improve CNN-based pneumonia detection by refining both the feature space and the model configuration. The framework may support automated chest X-ray screening on the studied dataset; however, external validation, statistical testing, and prospective clinical assessment remain necessary before any clinical deployment claim can be made.

Keywords: Pneumonia Detection ▪ Chest X-Ray Classification ▪ Convolutional Neural Networks ▪ Feature Selection ▪ Hyperparameter Optimization

1. INTRODUCTION

Pneumonia remains one of the most consequential respiratory infections worldwide, not only because of its immediate clinical severity but also because of the persistent burden it places on healthcare systems, families, and diagnostic services. The disease is commonly understood as an acute infection of the lung parenchyma in which the alveolar spaces and surrounding tissues become inflamed, often leading to fluid accumulation, impaired gas exchange, fever, cough, respiratory distress, and, in severe cases, hypoxemia or systemic deterioration. Although pneumonia affects individuals across all age groups, its impact is especially pronounced among children, whose immune systems, airway anatomy, and vulnerability to respiratory compromise make early recognition clinically important. The World Health Organization identifies pneumonia as a leading infectious cause of mortality in children, with the disease accounting for a substantial proportion of deaths among children under five years of age, particularly in resource-constrained regions where access to timely diagnosis and treatment may be limited [1]. Early diagnosis is therefore not a peripheral concern; it is a central requirement for reducing disease progression, preventing avoidable complications, and guiding appropriate clinical intervention.

Chest radiography has long occupied a critical position in pneumonia assessment because it provides a rapid and relatively accessible means of visualizing pulmonary abnormalities. Compared with advanced imaging modalities such as computed tomography, chest X-ray imaging is less expensive, more widely available, and more practical for routine diagnostic workflows, especially in high-volume clinical settings. In pediatric pneumonia, radiographs can support diagnostic evaluation by revealing patterns such as consolidation, infiltrates, and other opacity-related abnormalities that may not be sufficiently characterized through clinical examination alone. At the same time, radiographic diagnosis is not trivial. Interpretation can vary across readers, disease presentations, acquisition conditions, and image quality. This variability has motivated standardized approaches for pediatric chest radiograph interpretation, particularly in epidemiological and clinical research settings where consistent labeling is essential for reliable analysis [2, 3]. The diagnostic value of chest radiography, then, lies in its clinical accessibility and visual informativeness, but its practical use still depends heavily on expert interpretation and consistent assessment.

1.1 Clinical and Computational Background

Chest X-ray imaging plays a central role in pneumonia diagnosis because many radiographic manifestations of the disease are expressed through changes in pulmonary opacity, lung-field texture, and regional density distribution. In normal chest X-ray images, the lung regions usually appear relatively clear, with visible vascular markings and no substantial abnormal opacification. By contrast, pneumonia cases often show irregular patterns caused by inflammatory processes inside the lung parenchyma. Bacterial pneumonia may present as focal or lobar consolidation, whereas viral pneumonia may produce a more diffuse interstitial appearance distributed across broader pulmonary regions. These patterns are clinically meaningful, but they are also imperfectly separable. Their appearance may depend on disease stage, patient age, projection quality, comorbid conditions, and the subjective threshold used by the interpreting physician. Studies on pediatric chest-radiograph interpretation have therefore emphasized the need for standardized visual criteria and expert review when chest X-rays are used for pneumonia diagnosis or dataset construction [2, 3].

Despite the diagnostic value of chest radiography, manual interpretation remains dependent on radiological expertise, clinical context, and reader experience. This dependence becomes particularly challenging in healthcare environments where imaging volumes are high, radiologists are limited, or rapid triage decisions are required. The increasing availability of medical imaging data has consequently encouraged the development of automated computer-aided diagnostic systems that can support clinical decision-making by providing consistent preliminary assessments. Such systems are not intended to replace expert judgment. Their more defensible role is to assist screening, reduce workload, highlight cases requiring attention, and improve diagnostic efficiency when clinicians must process large numbers of radiographs. Large-scale chest X-ray datasets and deep-learning studies have shown that radiographic images contain learnable signals for thoracic disease recognition, including pneumonia-related patterns [4, 5]. This evidence has strengthened interest in automated chest X-ray analysis as a clinically relevant machine-learning application.

The present study employs the publicly available Chest X-Ray Images (Pneumonia) dataset obtained from Kaggle, which contains pediatric anterior-posterior chest X-ray images collected from Guangzhou Women and Children's Medical Center in Guangzhou, China [6]. The dataset is associated with the study by Kermany *et al.*, who investigated image-based deep learning for medical diagnosis and included pediatric chest radiographs as part of a broader evaluation of deep-learning-based disease identification [7]. According to the dataset documentation and the original study description, the radiographs were selected from retrospective pediatric cohorts, low-quality or unreadable images were excluded, and expert physicians reviewed diagnostic labels to improve annotation reliability [7, 6]. The dataset contains chest X-ray images categorized into Normal and Pneumonia classes, which makes it suitable for binary medical image classification. This binary formulation is methodologically useful because it provides a clear supervised-learning objective. That said, it also narrows the diagnostic scope: the task is pneumonia detection against normal cases, not multi-class differentiation among bacterial pneumonia, viral pneumonia, and other thoracic abnormalities.

In recent years, artificial intelligence and machine-learning approaches have demonstrated considerable potential in medical image analysis, particularly as imaging datasets and computational resources have expanded. Among these approaches, deep learning has become especially prominent because it can learn hierarchical feature representations directly from raw image data rather than relying exclusively on handcrafted descriptors. Convolutional Neural Networks (CNNs) are among the most widely adopted architectures for image classification because their convolutional filters are well suited to local feature extraction, spatial pattern recognition, and progressive representation learning. In medical imaging, CNNs have been applied to classification, detection, segmentation, registration, and related tasks, with survey evidence showing their broad adoption across several imaging modalities [8]. Their relevance to chest X-ray analysis is particularly strong because radiographic abnormalities are spatially structured: local opacities, edge patterns, and tissue-density variations must be interpreted in relation to anatomical context. CNN-based models can learn these local and hierarchical relationships more naturally than many conventional vector-based classifiers [9].

1.2 Motivation

Although CNNs have achieved promising performance across many image-classification applications, their effectiveness remains strongly dependent on several interrelated factors. The quality and relevance of extracted features directly influence the model's ability to distinguish between normal and pathological patterns. A CNN may learn discriminative pulmonary representations when the training data are informative and the model configuration is appropriate, but it may also learn unstable or incidental patterns when the dataset is limited, imbalanced, noisy, or insufficiently diverse. Architectural choices such as the number of convolutional layers, the number of filters, kernel size, activation functions, pooling operations, dense-layer width, and regularization strategy can substantially affect the learned representation. Hyperparameter settings are equally important. An unsuitable learning rate can slow convergence or destabilize training; an inappropriate batch size can affect gradient behavior; insufficient regularization can increase overfitting; and excessive model capacity can cause the network to memorize dataset-specific artifacts rather than learn robust radiographic indicators. These concerns are well recognized in deep medical image analysis, where dataset quality, transferability, and validation design strongly shape reported performance [8, 9].

Another important challenge arises from the presence of redundant or weakly informative representations within extracted feature spaces. Deep feature vectors can contain many dimensions, not all of which contribute equally to classification. Some features may encode overlapping visual patterns, while others may correspond to background structures, acquisition-specific variation, or activations that are only weakly related to pneumonia detection. High-dimensional representations may increase computational complexity and may also introduce unnecessary information that degrades classifier generalization. Feature-selection techniques address this issue by attempting to preserve informative variables while removing irrelevant or redundant ones. More specifically, feature selection can reduce dimensionality, improve model efficiency, simplify downstream classification, and potentially reduce overfitting by limiting the influence of noisy or unstable features [10]. In the context of CNN-based chest X-ray classification, feature selection is not a substitute for representation learning; rather, it can serve as a refinement mechanism applied to learned deep representations.

In addition to feature selection, hyperparameter optimization represents another critical aspect of CNN-based medical image analysis. Conventional manual hyperparameter tuning is often computationally expensive, subjective, and dependent on repeated trial-and-error experiments. Grid search can be inefficient when the search space is large, while random search may require many evaluations before locating competitive configurations. Automated hyperparameter optimization methods provide a more systematic alternative by exploring candidate configurations according to explicit performance criteria [11, 12]. Metaheuristic optimization algorithms are particularly attractive in this setting because they can search complex, nonlinear, mixed discrete-continuous spaces without requiring gradient information for the outer optimization problem. Algorithms such as Grey Wolf Optimizer, Particle Swarm Optimization, Whale Optimization Algorithm, Multi-Verse Optimizer, Satin Bowerbird Optimizer, and Genetic Algorithm have been widely used or adapted for difficult optimization tasks, including feature selection and model configuration [13, 14, 15, 16, 17, 18, 19]. The Ocotillo Optimization Algorithm (OcOA), a more recent desert-inspired metaheuristic, has also been introduced as an adaptive optimization method designed to balance exploration and exploitation in complex search spaces [20].

The integration of feature selection and hyperparameter optimization within a unified CNN-based framework therefore represents a meaningful research direction for medical image classification. Feature selection addresses the representation space by identifying a compact subset of informative features, whereas hyperparameter optimization addresses the model-configuration space by searching for training and architectural settings that improve learning behavior. These two forms of optimization are related but not identical. A good feature subset may reduce redundancy and improve classifier efficiency, while a good CNN configuration may improve convergence, regularization, and decision-boundary formation. Combining the two can therefore strengthen automated pneumonia detection by refining both what the model uses and how the model learns. This is the central motivation of the present work.

1.3 Research Problem

The primary research problem addressed in this study can be formulated as follows:

> How can the Ocotillo Optimization Algorithm (OcOA) be utilized to improve CNN-based pneumonia detection through feature selection and hyperparameter optimization?

This problem involves two interconnected objectives. The first objective concerns the selection of informative features capable of improving classification quality while reducing unnecessary feature dimensionality. The second objective focuses on identifying suitable CNN hyperparameter configurations that enhance training stability and classification effectiveness. Addressing these objectives jointly may improve the robustness and efficiency of automated pneumonia-detection systems. The term robustness is used here in an experimental sense: it refers to improved classification behavior under the evaluated dataset and protocol, not to unrestricted clinical generalization across hospitals, imaging devices, patient populations, or disease subtypes. That distinction is important because public chest X-ray benchmarks can support methodological evaluation, but external validation remains necessary before deployment-oriented conclusions can be made [7, 8].

1.4 Research Gap

Previous studies have extensively investigated CNN-based medical image classification, and several works have demonstrated the feasibility of deep-learning-based pneumonia detection from chest radiographs [4, 5, 7]. Similarly, feature selection and hyperparameter optimization have been studied as important components of machine-learning model development [10, 12]. Population-based metaheuristic algorithms have also been introduced independently and applied across a broad range of optimization problems, including binary feature selection and model-parameter search [14, 21, 13, 22, 15, 16, 17, 18, 19]. Likewise, the Ocotillo Optimization Algorithm has already been proposed as a standalone metaheuristic optimization method [20]. Consequently, the present work does not claim methodological novelty through the independent introduction of CNNs, feature selection, hyperparameter optimization, or OcOA as isolated concepts.

Instead, the research gap addressed in this study lies in the empirical assessment of an OcOA-guided CNN framework for pneumonia classification from chest X-ray images. More specifically, limited attention has been directed toward evaluating OcOA within a unified framework that simultaneously incorporates feature-selection analysis, post-feature-selection machine-learning baseline comparison, and optimizer-guided CNN hyperparameter optimization. This combined evaluation is important because an optimizer may behave differently depending on whether it is used to select binary feature subsets or to tune CNN hyperparameters. A method that selects compact feature subsets does not automatically guarantee strong CNN configuration performance, and a method that tunes CNN hyperparameters effectively does not necessarily provide the best feature-selection behavior. Investigating these stages together provides a broader understanding of how OcOA can contribute to CNN-based medical image classification, particularly for pediatric chest X-ray pneumonia detection.

1.5 Contributions

The contributions of this study can be summarized as follows. First, a hybrid OcOA+CNN framework is evaluated for binary pneumonia classification using chest X-ray images derived from a widely used pediatric radiographic dataset [7, 6]. Second, OcOA is employed as a feature-selection mechanism and compared against several binary metaheuristic optimization algorithms, including binary Grey Wolf Optimizer (bGWO), binary Particle Swarm Optimization (bPSO), binary Stochastic Fractal Search (bSFS), binary Whale Optimization Algorithm (bWOA), binary Multi-Verse Optimizer (bMVO), binary Satin Bowerbird Optimizer (bSBO), and binary Genetic Algorithm (bGA) [13, 14, 21, 22, 15, 16, 17, 18, 19]. Third, the selected features are evaluated using CNN and conventional machine-learning baseline models to examine classification effectiveness after feature reduction. Fourth, OcOA is further utilized for CNN hyperparameter optimization in order to improve model configuration and training behavior [20, 12]. Finally, the optimized OcOA+CNN framework is compared with multiple optimizer-CNN variants using evaluation metrics including accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F-score. These metrics are reported because binary medical classification requires more than a single overall accuracy value; sensitivity and specificity, in particular, help describe the model's behavior with respect to pneumonia-case recognition and normal-case recognition.

1.6 Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews the related literature associated with pneumonia detection, CNN-based medical image analysis, feature selection, and metaheuristic optimization algorithms. Section 3 describes the dataset and its clinical characteristics. Section 4 presents the proposed methodology, including preprocessing, feature selection, and hyperparameter optimization procedures. Section 5 explains the experimental setup and evaluation criteria. Section 6 reports the obtained experimental results. Section 7 discusses the findings and their implications. Section 8 outlines the limitations of the study. Finally, Section 9 concludes the paper and highlights potential future research directions.

2. RELATED WORK

Chest X-ray-based pneumonia detection has become a recurrent benchmark problem in medical-image analysis because it combines a clinically meaningful diagnostic objective with an imaging modality that is comparatively accessible, low-cost, and routinely used in respiratory assessment. Pneumonia can manifest in radiographs through heterogeneous opacity patterns, consolidation, and interstitial changes, yet these findings are not always visually obvious, particularly when image quality varies or when the disease presentation is subtle. This diagnostic ambiguity has encouraged sustained interest in computational screening systems that can learn discriminative patterns directly from radiographic data and assist clinicians by providing rapid preliminary classification. Within this setting, deep learning has been especially influential, not because it removes the need for clinical interpretation, but because it can encode image patterns at a scale and consistency that are difficult to maintain under high workload conditions.

Early large-scale chest X-ray studies helped establish the technical foundation for automated thoracic disease analysis. Wang et al. introduced ChestX-ray8, a hospital-scale database containing more than one hundred thousand frontal chest radiographs with weakly supervised thoracic disease labels, thereby demonstrating that large radiographic repositories could support deep-learning-based classification and localization of common thoracic abnormalities [4]. Although ChestX-ray8 was designed as a multi-label thoracic disease benchmark rather than a binary pediatric pneumonia dataset, it showed that convolutional representations could be trained on chest radiographs and adapted to clinically relevant recognition tasks. Rajpurkar et al. later presented CheXNet, a DenseNet-based model trained for pneumonia detection on chest X-rays, and their work became widely cited because it framed pneumonia recognition as a high-impact radiographic classification problem and emphasized the diagnostic potential of CNN-based feature learning [5]. These studies helped move chest X-ray analysis from hand-engineered radiographic descriptors toward end-to-end deep visual representation learning.

The dataset used in the present study follows a more focused binary formulation. It is derived from pediatric anterior-posterior chest radiographs associated with the work of Kermany et al., who investigated image-based deep learning for medical diagnosis and included pediatric chest X-ray images collected from Guangzhou Women and Children's Medical Center [7]. The widely used Kaggle release of this resource, commonly referenced as the Chest X-Ray Images (Pneumonia) dataset, contains chest radiographs organized into Normal and Pneumonia categories and has become a frequent benchmark for binary pneumonia classification [6]. This Normal-versus-Pneumonia structure is attractive for method development because it provides a direct diagnostic endpoint and allows researchers to compare image classifiers under a clear supervised-learning setting. That said, the apparent simplicity of the binary task should not be overstated. The images originate from a pediatric cohort, the class distributions are not perfectly balanced, and the radiographic appearance of pneumonia may vary across infection type, disease stage, acquisition condition, and patient anatomy. A model that performs strongly on this dataset therefore provides useful experimental evidence, but not by itself a complete claim of clinical generalization.

2.1 Chest X-Ray-Based Pneumonia Detection

Chest X-ray-based pneumonia detection studies generally formulate the problem in one of two ways. The first formulation treats pneumonia as one label within a broader multi-label thoracic abnormality recognition task, as in large chest radiography datasets where pneumonia may coexist with other findings such as atelectasis, effusion, or infiltration [4]. This formulation is clinically rich, but methodologically difficult, because labels may be noisy, overlapping, and derived from reports rather than direct expert image annotation. The second formulation treats pneumonia detection as a binary classification problem, where each image is assigned either to a Normal class or a Pneumonia class. The present paper follows the second formulation, which is common in studies using the pediatric Chest X-Ray Images (Pneumonia) dataset [7, 6]. Binary classification reduces the output space and allows the methodological focus to shift toward representation quality, feature selection, classifier behavior, and optimization strategy.

The popularity of the pediatric pneumonia dataset is partly due to its compact but clinically grounded structure. The radiographs were collected from pediatric patients and labeled according to the presence or absence of pneumonia, which makes the dataset suitable for evaluating supervised image-classification pipelines in a controlled setting [7]. Unlike broad thoracic datasets that include many disease labels and uncertain co-occurrences, this dataset supports a direct diagnostic distinction. It has therefore been repeatedly used to evaluate CNNs, transfer-learning architectures, ensemble classifiers, and hybrid machine-learning frameworks. The same property also makes it suitable for the present work, where the central question is not whether pneumonia can be represented visually in chest radiographs, but whether an optimization-guided CNN framework can improve classification behavior through feature selection and hyperparameter tuning.

A recurring observation in this research area is that high classification accuracy on a single public dataset must be interpreted carefully. Chest radiographs are sensitive to acquisition protocol, patient age, scanner characteristics, institutional imaging practice, and label verification procedure. A model trained and evaluated on pediatric images from one source may not behave identically on adult radiographs, portable intensive-care images, or data acquired from a different hospital. Kermany et al. explicitly positioned deep learning as a diagnostic-support technology rather than a replacement for clinical judgment, which is an important distinction for studies that report high accuracy on curated datasets [7]. The same caution applies here. The related literature shows strong evidence that chest X-rays contain learnable signals for pneumonia classification, but it also shows that dataset provenance and validation design are central to the credibility of any proposed method.

2.2 CNNs in Medical Image Classification

Convolutional Neural Networks have become central to medical image classification because their inductive structure matches the spatial nature of imaging data. A chest X-ray is not a simple vector of independent measurements; it is a two-dimensional projection in which local intensity patterns, edges, textures, and anatomical structures interact across the lung fields. CNNs exploit this structure through convolutional filters that scan local receptive fields and learn repeated visual motifs. In early layers, these filters often respond to low-level features such as edges, gradients, and small texture changes. In deeper layers, the network can combine these local activations into more abstract patterns associated with anatomical boundaries, diffuse opacity, focal consolidation, or other radiographic signals. This hierarchical pattern learning is one reason CNNs have been widely adopted in medical-image analysis [8, 9].

The value of CNNs in chest X-ray classification is also connected to spatial representation. Pneumonia-related abnormalities may be localized to a particular lobe, appear as asymmetric opacity, or present as broader interstitial texture changes. A conventional classifier trained only on flattened pixels would struggle to preserve the spatial continuity of these findings. By contrast, convolution and pooling operations allow CNNs to retain local neighborhood information while progressively building translation-tolerant representations. This property is particularly useful in chest radiographs because the exact location, size, and intensity of pathological patterns may differ across patients. CNNs can therefore learn discriminative representations without requiring handcrafted feature definitions for each possible radiographic manifestation.

Medical-image classification studies have repeatedly shown that CNN performance depends not only on the architecture but also on dataset characteristics, transfer-learning strategy, input resolution, and training configuration [9]. This dependency is important for the present study because it motivates the use of hyperparameter optimization. A CNN with an unsuitable learning rate may converge slowly or become unstable. A network with insufficient representational capacity may underfit subtle radiographic patterns, whereas an excessively large model may memorize training examples, especially when the dataset is limited. Dropout rate, batch size, optimizer choice, number of filters, kernel size, and dense-layer width can all change the bias–variance behavior of the model. There is no universal configuration that is optimal across all medical imaging datasets.

Overfitting remains a central concern. Medical datasets are often smaller than natural-image datasets, and labels may contain uncertainty due to inter-reader variability or imperfect annotation processes. Even when the number of images appears adequate, the effective diversity of disease presentation may be limited by patient demographics, acquisition site, and imaging protocol. CNNs can learn highly predictive but clinically incidental correlations if the training data contain confounding signals. This issue is not unique to pneumonia detection, but it is especially relevant in radiographic classification because image acquisition artifacts, positioning differences, and preprocessing choices can influence the learned feature space. Deep-learning surveys in medical imaging have therefore emphasized the importance of validation design, data diversity, and interpretability when assessing CNN-based systems [8].

Computational cost is another practical limitation. CNN training requires repeated forward and backward passes through large image tensors, and hyperparameter search multiplies this cost because several candidate configurations must be trained or partially trained. Even after training, inference time and memory footprint matter when a model is intended for deployment in screening workflows. A highly accurate model that is computationally inefficient may be less useful in resource-constrained settings. These limitations do not weaken the case for CNNs; rather, they clarify why CNNs benefit from systematic optimization. The present paper builds on this view by examining whether OcOA can improve a CNN-based pneumonia classifier at two levels: the feature space and the hyperparameter configuration.

2.3 Feature Selection for Medical Image Models

Feature selection is a dimensionality-reduction strategy that aims to retain informative variables while removing redundant, irrelevant, or weakly discriminative ones. In classical machine learning, feature selection is often used to reduce model complexity, improve interpretability, and limit overfitting. In deep-learning-based medical-image analysis, the role of feature selection is more nuanced. CNNs already perform a form of representation learning, yet the deep features extracted from intermediate or late network layers can still be high-dimensional and partially redundant. Some activations may encode overlapping visual patterns, while others may reflect background variation, acquisition artifacts, or features that do not contribute meaningfully to the diagnostic decision. Selecting a compact subset of these representations can therefore support more efficient classification.

General feature-selection literature distinguishes among filter, wrapper, and embedded approaches [10]. Filter methods evaluate features using statistical criteria independent of a final classifier. Embedded methods perform feature selection as part of model training. Wrapper methods, by contrast, evaluate candidate feature subsets according to classifier performance and are often more flexible, though computationally more expensive. Metaheuristic feature selection usually belongs to the wrapper family because each candidate solution represents a subset of features and is evaluated through a fitness function. This structure is suitable for medical-image features because the relationship between feature subsets and classification performance is rarely linear or additive. A feature may be weak alone but useful in combination with others.

For medical image models, feature selection can serve several practical purposes. First, it can reduce redundant features, which is important when deep representations contain correlated activations produced by adjacent filters or similar learned patterns. Second, it can improve classification efficiency by lowering the dimensionality of the input representation passed to downstream classifiers. Third, it may reduce inference complexity when the selected feature subset allows a classifier to operate on a smaller representation. This point should be treated carefully: feature selection does not automatically guarantee lower end-to-end inference time unless the implementation actually avoids computing unused features or reduces downstream computational cost. Still, at the classifier-input level, a smaller feature vector is usually easier to process than a larger one. Fourth, feature selection can potentially improve generalization by discouraging the classifier from relying on noisy or unstable representations.

In pneumonia detection from chest X-rays, feature selection is particularly relevant because the discriminative information may be concentrated in a subset of learned radiographic patterns. Normal and pneumonia images differ not only in global intensity distribution but also in localized and texture-level characteristics. If a CNN-derived feature vector contains many dimensions that do not correspond to these diagnostic differences, a downstream classifier may become unnecessarily complex. The present paper therefore uses feature selection not as a replacement for CNN representation learning, but as a refinement step applied after deep feature extraction. This design recognizes that CNNs are effective feature learners while also acknowledging that their learned representations may still benefit from optimization-guided pruning.

2.4 Metaheuristic Optimization for Feature Selection

Metaheuristic optimization algorithms are frequently used for feature selection because the search space is combinatorial. For a feature vector with d candidate dimensions, there are 2^d possible subsets. Exhaustive search is therefore infeasible except for very small feature spaces. Binary metaheuristic algorithms address this difficulty by representing each candidate solution as a binary vector, where a value of one denotes a selected feature and a value of zero denotes an excluded feature. The optimizer then searches for a subset that balances predictive performance with dimensionality reduction. This structure is especially suitable for wrapper-based feature selection, where the objective function can include both classification error and selected-feature ratio.

The present study compares bOcOA with several established binary metaheuristic algorithms. OcOA is a recently proposed desert-inspired metaheuristic designed to balance exploration and exploitation through adaptive search behavior [20]. In its binary form, bOcOA can be adapted to feature selection by converting continuous search positions into binary decisions through a transfer mechanism. This adaptation allows the algorithm to explore candidate subsets rather than continuous-valued parameter vectors. Because the present work evaluates bOcOA empirically against other binary optimizers, its contribution lies in assessing whether OcOA can provide useful feature-selection behavior for CNN-derived chest X-ray representations.

The comparison set includes bGWO, bPSO, bSFS, bWOA, bMVO, bSBO, and bGA. GWO models leadership hierarchy and hunting behavior in grey wolves and has been widely adopted because of its relatively simple structure and exploration–exploitation balance [13]. PSO originates from swarm intelligence and updates candidate solutions according to individual and collective experience [14]. Its discrete binary version was introduced to handle binary-valued optimization problems, making it directly relevant to feature subset selection [21]. SFS is a mathematics-inspired metaheuristic based on fractal diffusion processes and stochastic search mechanisms [22]. WOA models the bubble-net hunting strategy of humpback whales and has been widely applied to continuous and adapted binary optimization tasks [15]. MVO draws inspiration from cosmological concepts such as white holes, black holes, and wormholes to model global search and solution refinement [16]. SBO is inspired by the mating behavior of satin bowerbirds and has been used as an optimizer in machine-learning model configuration and parameter estimation tasks [17]. GA, grounded in evolutionary computation, uses selection, crossover, and mutation to evolve candidate solutions and remains one of the most established population-based search methods [18, 19].

In the present manuscript, the reported feature-selection comparison uses bWOA, denoting binary Whale Optimization Algorithm. The abbreviation bWAO is not adopted because the experimental table corresponds to bWOA rather than a separately confirmed algorithm with the abbreviation bWAO. This distinction is necessary for terminological consistency and for preserving alignment between the related-work discussion and the reported experiments. The broader methodological point remains the same: binary metaheuristic feature selection provides a flexible mechanism for exploring high-dimensional CNN feature spaces without requiring differentiability of the feature-subset objective.

The advantage of metaheuristic feature selection is not that it guarantees a globally optimal subset. It does not. Its practical value lies in its ability to search large, irregular, nonconvex subset spaces where exact optimization is impractical. In medical-image classification, this flexibility can be useful because the selected feature subset must often balance several objectives at once: low classification error, compact representation size, robustness across stochastic runs, and manageable computational demand. The present paper situates bOcOA within this family of methods and evaluates whether its search behavior is competitive when applied to deep features extracted from chest X-ray images.

2.5 Metaheuristic Optimization for CNN Hyperparameters

CNN hyperparameter optimization is a difficult problem because training behavior depends on interacting architectural and learning parameters. Learning rate affects the scale of gradient updates and can determine whether the model converges smoothly, oscillates, or stagnates. Batch size influences gradient noise, memory consumption, and training stability. The number of convolutional filters controls representational capacity, while kernel size affects the local spatial patterns captured by each convolutional layer. Dense units, dropout rate, optimizer choice, and number of epochs further shape the model's ability to fit the training data without losing generalization. These parameters do not act independently. A larger network may require stronger regularization; a smaller batch size may interact with learning rate; a longer training schedule may improve convergence but also increase overfitting risk.

Manual tuning is therefore limited. It depends heavily on trial-and-error, prior experience, and computational patience. Grid search can become inefficient because it evaluates many unpromising combinations, while random search improves coverage but still lacks adaptive feedback from previous evaluations. Metaheuristic optimization offers an alternative by treating hyperparameter selection as a search problem in which each candidate solution encodes a CNN configuration. The candidate is evaluated according to validation or training-performance criteria, and the optimizer updates the population to explore more promising regions of the search space. This approach is computationally expensive because it may require training many candidate models, but it can be effective when the search space is mixed, nonlinear, and partly discrete.

The present paper evaluates OcOA+CNN against GWO+CNN, GA+CNN, PSO+CNN, and WOA+CNN. These optimizer-CNN variants are grounded in the same metaheuristic families discussed above, but the search object changes from a binary feature subset to a CNN hyperparameter configuration. OcOA+CNN uses the Ocotillo Optimization Algorithm to identify a suitable CNN configuration [20]. GWO+CNN uses Grey Wolf Optimizer principles for hyperparameter search [13]. GA+CNN relies on evolutionary operators to recombine and mutate candidate model configurations [18, 19]. PSO+CNN uses swarm-based movement toward promising hyperparameter regions [14]. WOA+CNN uses the encircling and spiral-search mechanisms associated with the Whale Optimization Algorithm [15]. In the reported optimizer-CNN comparison, the correct abbreviation is PSO+CNN rather than POS+CNN, because the corresponding optimizer is Particle Swarm Optimization.

For CNN-based pneumonia detection, hyperparameter optimization is not a cosmetic step. A model may fail to generalize because its architecture is too shallow, because its dropout rate is poorly matched to the dataset, or because the learning rate prevents stable convergence. In medical imaging, where datasets are often limited and class distributions may be imbalanced, these choices can materially affect sensitivity and specificity. This is clinically important because a false negative may correspond to a missed pneumonia case, whereas a false positive may lead to unnecessary follow-up. Optimizing a CNN therefore requires attention not only to overall accuracy but also to class-specific metrics such as sensitivity, specificity, PPV, NPV, and F-score.

The related literature supports the general use of CNNs for medical-image analysis and the use of metaheuristic algorithms for difficult optimization problems [8, 9, 13, 14, 15]. However, the value of a specific optimizer must still be evaluated in the target task. An algorithm that performs well on mathematical benchmark functions may not necessarily produce the best CNN configuration for pediatric chest X-ray classification. The present paper therefore treats OcOA+CNN as an empirical question: whether OcOA, when applied to CNN hyperparameter tuning, improves pneumonia-detection performance under the dataset and evaluation protocol used in this study.

The reviewed studies cover three methodological streams that motivate the present work: chest X-ray pneumonia detection, convolutional representation learning, and metaheuristic optimization for feature selection and hyperparameter tuning. Table 1 summarizes the main studies and algorithms discussed in this section, with emphasis on their role in shaping the proposed Ocotillo Optimization Algorithm-guided Convolutional Neural Network framework.

2.6 Summary of Literature Gap

The reviewed literature shows that CNNs are well suited to medical-image classification because they learn local and hierarchical visual representations directly from image data [8, 9]. Chest X-ray studies have further shown that radiographic images contain learnable signals for pneumonia and thoracic disease recognition [4, 5]. The pediatric Chest X-Ray Images (Pneumonia) dataset has become a widely used benchmark for Normal-versus-Pneumonia classification because it provides a focused binary diagnostic task derived

Table 1. Summary of the main studies and optimization methods discussed in the related work.

Reference	Main focus	Technical contribution	Relevance to the present study
[4]	Large-scale thoracic disease recognition	Introduced ChestX-ray8 as a large frontal chest radiography dataset with weakly supervised disease labels.	Established the value of large chest X-ray repositories for training deep-learning models on thoracic abnormalities.
[5]	Pneumonia detection from chest radiographs	Presented CheXNet, a DenseNet-based Convolutional Neural Network model for pneumonia recognition.	Demonstrated the diagnostic potential of deep convolutional representations for pneumonia classification.
[7]	Deep learning for medical diagnosis	Used pediatric chest X-ray images to investigate image-based diagnosis using deep learning.	Provides the clinical and dataset foundation for the Normal-versus-Pneumonia classification task used in this study.
[6]	Public pneumonia dataset release	Provided the Kaggle version of the Chest X-Ray Images (Pneumonia) dataset.	Supplies the benchmark dataset used for evaluating the proposed framework.
[8]	Deep learning in medical image analysis	Reviewed deep-learning methods across medical imaging tasks and highlighted validation and generalization issues.	Supports the methodological motivation for careful evaluation of Convolutional Neural Network-based medical classifiers.
[9]	Deep convolutional learning in medical imaging	Examined deep Convolutional Neural Network architectures and transfer learning for computer-aided diagnosis.	Motivates the use of convolutional representation learning for radiographic image classification.
[10]	Feature-selection methods	Reviewed filter, wrapper, and embedded feature-selection strategies.	Provides the conceptual basis for using wrapper-style metaheuristic feature selection over extracted features.
[20]	Ocotillo Optimization Algorithm	Introduced the Ocotillo Optimization Algorithm as a population-based metaheuristic optimizer.	Provides the optimization method used in this paper for feature selection and Convolutional Neural Network hyperparameter tuning.
[13]	Grey Wolf Optimizer	Proposed Grey Wolf Optimizer as a population-based method inspired by grey wolf leadership and hunting behavior.	Serves as one of the competing optimizers in the feature-selection and optimizer-Convolutional Neural Network comparisons.
[14]	Particle Swarm Optimization	Introduced Particle Swarm Optimization as a swarm-intelligence algorithm based on social learning behavior.	Provides the foundation for Particle Swarm Optimization-based comparison models.
[21]	Binary Particle Swarm Optimization	Adapted Particle Swarm Optimization to binary-valued optimization problems.	Supports the binary feature-selection formulation used for selected-feature encoding.
[22]	Stochastic Fractal Search	Proposed Stochastic Fractal Search using fractal diffusion and stochastic update mechanisms.	Included as a competing binary metaheuristic method for feature selection.
[15]	Whale Optimization Algorithm	Introduced Whale Optimization Algorithm based on humpback whale bubble-net hunting behavior.	Used as a competing optimizer in both binary feature selection and Convolutional Neural Network hyperparameter optimization.
[16]	Multiverse Optimization	Proposed Multiverse Optimization using cosmological search concepts such as white holes, black holes, and wormholes.	Included as a competing binary optimizer in the feature-selection stage.
[17]	Satin Bowerbird Optimizer	Introduced Satin Bowerbird Optimizer as a population-based search method inspired by bowerbird mating behavior.	Included as a competing binary optimizer for feature-selection comparison.
[18]; [19]	Genetic Algorithm	Established evolutionary optimization using selection, crossover, and mutation.	Provides the foundation for the Genetic Algorithm-based feature-selection and optimizer-Convolutional Neural Network comparisons.

from clinically collected pediatric radiographs [7, 6]. At the same time, CNN performance remains sensitive to data quality, model configuration, overfitting behavior, and validation design.

Existing work also shows that feature selection and metaheuristic optimization can be useful for high-dimensional machine-learning problems [10]. Binary metaheuristic algorithms provide a practical way to search large feature-subset spaces, while population-based optimizers can be adapted to CNN hyperparameter tuning. Yet the combined use of OcOA in both stages has received limited task-specific evaluation for CNN-based pneumonia detection from chest X-rays. This is the gap addressed by the present paper. Specifically, the study evaluates whether OcOA can improve a CNN-based pneumonia-detection framework when used first for feature selection over CNN-derived representations and then for CNN hyperparameter optimization. The contribution is therefore not the independent invention of CNNs, feature selection, or metaheuristic optimization, but a focused empirical assessment of an OcOA-guided two-stage optimization pipeline for pediatric chest X-ray pneumonia classification.

3. DATASET DESCRIPTION

The dataset used in this study is the publicly available *Chest X-Ray Images (Pneumonia)* dataset, which has been widely used as a benchmark resource for automated pneumonia detection from radiographic images. The dataset is appropriate for the present work because it provides a binary classification setting in which chest X-ray images are categorized as either normal or pneumonia cases. This structure aligns directly with the objective of evaluating a convolutional neural network-based diagnostic framework supported by feature selection and hyperparameter optimization.

3.1 Dataset Source

The *Chest X-Ray Images (Pneumonia)* dataset was obtained from Kaggle and consists of 5,863 chest X-ray images stored in JPEG format. The dataset contains two diagnostic categories, namely *Normal* and *Pneumonia*, and is organized into three main folders corresponding to the training, testing, and validation subsets. The original source of the dataset is Mendeley Data, where the dataset is made available under the Creative Commons Attribution 4.0 International license. The Kaggle repository reports that the dataset contains 5,863 JPEG chest X-ray images distributed across the two categories and arranged into *train*, *test*, and *val* directories, while the corresponding Mendeley Data record identifies the dataset as openly available under the CC BY 4.0 license.

3.2 Dataset Origin and Clinical Context

The images in this dataset are anterior-posterior chest radiographs selected from retrospective pediatric cohorts. The patients were between one and five years of age, and the images were collected from Guangzhou Women and Children's Medical Center in Guangzhou, China. Since all radiographs were acquired as part of routine clinical care, the dataset reflects a clinically grounded imaging scenario rather than a purely synthetic or laboratory-controlled collection.

The dataset documentation describes a multi-stage quality-control and diagnostic verification process. First, low-quality or unreadable chest radiographs were removed to reduce noise and prevent diagnostically unreliable scans from entering the learning pipeline. The remaining images were then graded by two expert physicians to establish diagnostic labels. To further reduce labeling errors, the evaluation set was additionally reviewed by a third expert. This annotation process is important because supervised medical image classification depends strongly on label reliability; inaccurate labels can distort the learning objective and weaken the clinical validity of model evaluation.

3.3 Class Definitions

The dataset is used in this study as a binary classification dataset. The first class is *Normal*, which represents chest X-ray images without radiographic evidence of pneumonia. These images generally show clearer lung fields and do not exhibit the abnormal opacification patterns associated with pulmonary infection. The second class is *Pneumonia*, which represents chest X-ray images containing radiographic findings consistent with pneumonia.

Although the dataset documentation provides illustrative examples that distinguish bacterial and viral pneumonia patterns, these examples are treated only as clinical context in the present study. Bacterial pneumonia may appear as focal lobar consolidation, whereas viral pneumonia may show a more diffuse interstitial pattern. However, the experimental formulation in this work is restricted to binary classification between *Normal* and *Pneumonia*. Therefore, bacterial and viral pneumonia are not treated as separate output classes in the proposed classification framework.

3.4 Dataset Split

The dataset follows the original folder organization provided by the Kaggle repository. It is divided into three subsets: a training set, a testing set, and a validation set. The training subset is used for model learning, including the extraction of discriminative patterns from chest X-ray images. The testing subset is used to evaluate the final classification behavior of the trained models, while the validation subset provides an additional partition for monitoring model behavior during development and optimization.

Table 2 summarizes the dataset structure used in this study. The table reports only the information provided by the dataset documentation and avoids introducing unverified class-level image counts.

Table 2. Structure of the Chest X-Ray Images (Pneumonia) dataset.

Attribute	Description
Dataset name	Chest X-Ray Images (Pneumonia)
Source	Kaggle
Original source	Mendeley Data
Total number of images	5,863
Image format	JPEG
Classes	Normal and Pneumonia
Folder organization	train, test, val
License	CC BY 4.0

3.5 Dataset Figure

Visual inspection of chest X-ray images remains an important component of pneumonia assessment because radiographic abnormalities may appear with varying intensity, spatial distribution, and anatomical localization across patients. Although the proposed framework performs automated feature extraction and classification, examining representative samples from the dataset provides important clinical and methodological context regarding image variability, pulmonary appearance, and radiographic quality. The dataset used in this study contains pediatric anterior-posterior chest radiographs collected under routine clinical conditions, and the visual examples illustrate the diversity of lung-field appearance observed across the dataset.

The representative chest X-ray samples shown in Figure 1 demonstrate variations in thoracic structure, lung opacity distribution, rib visibility, and image contrast among pediatric radiographs. Several images exhibit relatively clear lung regions associated with normal radiographic appearance, whereas others contain visible opacity patterns and pulmonary-density changes consistent with pneumonia-related findings. Differences in image brightness, anatomical positioning, and acquisition characteristics are also observable across the samples. These variations highlight the practical challenges associated with automated chest X-ray analysis because the classification framework must distinguish clinically meaningful pulmonary abnormalities while remaining robust to acquisition-related variability.

The figure also demonstrates why convolutional neural networks are suitable for chest radiograph analysis. The local and distributed radiographic patterns associated with pneumonia may vary considerably between patients, and the relevant abnormalities are not confined to a single fixed spatial location. Consequently, CNN-based representation learning becomes valuable because convolutional operations can capture hierarchical texture and opacity patterns directly from the image data. The observed variability across the dataset fur-

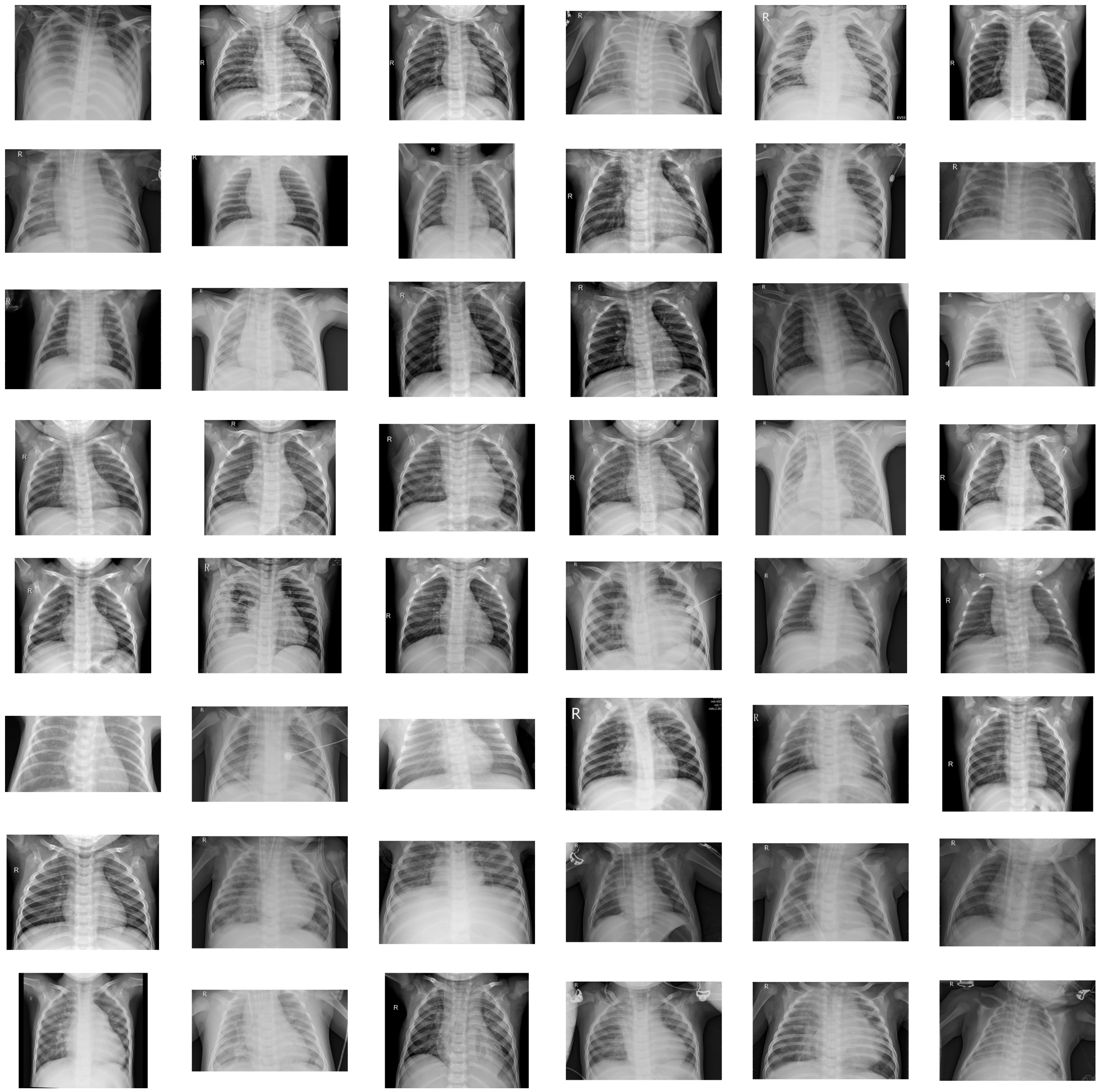

Figure 1. Representative chest X-ray samples from the Chest X-Ray Images (Pneumonia) dataset used in this study. The displayed pediatric anterior-posterior radiographs illustrate variations in lung appearance, pulmonary opacity distribution, image contrast, and anatomical structure across the dataset.

ther motivates the use of feature selection and hyperparameter optimization to improve the robustness and discriminative capability of the proposed classification framework.

4. MATERIALS AND METHODS

This section describes the methodological framework adopted for pneumonia classification from chest X-ray images. The proposed approach integrates convolutional neural network-based representation learning with Ocotillo Optimization Algorithm (OcOA)-based feature selection and hyperparameter optimization. The methodology is designed to ensure that the classification process is not treated as a single isolated training stage, but rather as a structured pipeline in which image preparation, feature representation, feature subset selection, model comparison, and final optimization are performed sequentially. This organization is important for reproducibility because each stage affects the behavior of the subsequent stage and, consequently, the final classification decision.

4.1 Overview of the Proposed Framework

The proposed framework begins with the input chest X-ray images obtained from the dataset described in Section 3. Each image is first passed through the preprocessing stage to ensure that the data are represented in a consistent numerical format suitable for computational analysis. After preprocessing, data analysis is performed to examine the input structure and support the subsequent modeling stages. The processed images are then used within a convolutional neural network (CNN) component, which provides image-based representation learning for the pneumonia-classification task.

Following CNN-based representation learning, OcOA is employed as a feature-selection mechanism. In this stage, the optimizer searches for a compact and informative subset of features by balancing classification-related quality with feature reduction. The selected features are then evaluated using both CNN-based and conventional machine-learning classifiers. This intermediate evaluation allows the study to examine whether the selected representations remain discriminative across different model families, rather than being useful only for a single classifier.

The final stage of the framework applies OcOA to CNN hyperparameter optimization. Instead of relying only on manually selected CNN configurations, OcOA is used to search for improved hyperparameter settings that support more effective training and classification. The optimized CNN model is then used for the final pneumonia-classification task. Model performance is evaluated using accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F-score. These metrics provide complementary information about the classifier's ability to identify pneumonia cases, recognize normal cases, and maintain balanced predictive behavior across the two diagnostic categories.

The complete methodological flow is organized as follows: input chest X-ray images are first preprocessed and analyzed; CNN-based representation learning is then performed; OcOA is applied for feature selection; the selected features are evaluated using machine-learning baseline models; OcOA is subsequently used for CNN hyperparameter optimization; and the final pneumonia classification is assessed using the selected evaluation metrics. This framework is intended to provide a systematic assessment of the role of OcOA in both feature-level and model-level optimization.

4.2 Data Preprocessing

Data preprocessing was conducted before feature selection and model training to ensure that all chest X-ray images were represented consistently and could be processed by the CNN architecture. Because chest X-ray images in public medical datasets may differ in spatial resolution, intensity distribution, and file-level characteristics, preprocessing is a necessary step for reducing unwanted variability that is unrelated to the diagnostic categories. In the proposed pipeline, preprocessing prepares the input data for CNN-based representation learning while preserving the radiographic patterns required for distinguishing normal and pneumonia cases.

The preprocessing stage includes image resizing, pixel-value normalization, and label encoding. Image resizing standardizes the spatial dimensions of the input images so that they are compatible with the CNN input layer. Pixel normalization transforms raw intensity values into a numerically stable range suitable for gradient-based learning, which can improve training stability and reduce sensitivity to scale differences across images. The class labels are encoded into a machine-readable format corresponding to the two categories used in this study, namely *Normal* and *Pneumonia*.

When augmentation is applied, it should be performed only on the training subset to increase sample variability without contaminating the validation or testing subsets. Augmentation operations such as rotation, zooming, horizontal flipping, and contrast adjustment may help expose the model to plausible image variations, but these operations must be selected carefully in medical imaging because excessive transformations can distort clinically meaningful structures. Similarly, any class-imbalance handling strategy must be applied within the training process only, so that validation and testing performance reflect the original evaluation distribution rather than an artificially modified one.

The preprocessing procedure is positioned before the feature-selection stage because OcOA receives feature representations derived from the prepared image data. Consequently, consistent preprocessing is essential for ensuring that the optimizer evaluates feature subsets under stable input conditions. It also supports fair comparison among the machine-learning baselines and optimizer-CNN models because all models operate on data produced by the same preprocessing pipeline.

4.3 CNN Architecture

The convolutional neural network (CNN) employed in this study was designed to perform binary pneumonia classification from chest X-ray images while maintaining a relatively balanced trade-off between representation-learning capability and computational complexity. The architecture was constructed to capture hierarchical spatial characteristics from radiographic images through a sequence of convolutional, pooling, and fully connected operations. Prior to hyperparameter optimization, the CNN was configured using a conventional baseline architecture suitable for medical image classification tasks.

The input images were resized to a spatial resolution of 224×224 pixels with three image channels, resulting in an input shape of $224 \times 224 \times 3$. This image size was selected because it preserves sufficient anatomical detail while maintaining computational feasibility during model training and optimization. The network architecture consisted of four convolutional layers organized sequentially to enable progressive extraction of low-level and high-level image representations.

The first convolutional layer employed 32 filters with a kernel size of 3×3, followed by a Rectified Linear Unit (ReLU) activation function. The second convolutional layer used 64 filters with the same 3×3 kernel configuration and ReLU activation. The third and fourth convolutional layers contained 128 and 256 filters, respectively, also using 3×3 kernels and ReLU activation functions. ReLU activation was selected because of its computational efficiency and its ability to reduce vanishing-gradient behavior during deep-network training.

To reduce spatial dimensionality and preserve the most informative feature responses, max-pooling layers with a pooling size of 2×2 were inserted after convolutional operations. Max pooling assists in reducing feature-map dimensionality while maintaining dominant activation patterns associated with pulmonary abnormalities in chest X-ray images. This dimensionality reduction also contributes to lowering computational overhead during subsequent learning stages.

Following the convolutional and pooling operations, the extracted feature maps were flattened into a one-dimensional representation before entering the fully connected layers. The dense component of the network consisted of two fully connected layers containing 512 and 256 neurons, respectively. ReLU activation was also used within the dense layers to maintain nonlinear representation capability throughout the network.

To reduce the risk of overfitting, a dropout layer with a dropout rate of 0.5 was applied between the dense layers.

Dropout regularization randomly deactivates a subset of neurons during training, thereby encouraging the network to learn more generalized feature representations rather than overly memorizing the training data. This is particularly important in medical image analysis because model generalization to previously unseen radiographs is essential for reliable classification performance.

The output layer consisted of a single neuron configured for binary classification. A sigmoid activation function was used in the final layer to generate probabilistic predictions corresponding to the two diagnostic categories, namely *Normal* and *Pneumonia*. The sigmoid function was selected because it is widely used for binary medical-classification tasks and produces interpretable output probabilities within the interval $[0, 1]$.

The CNN architecture used in this study therefore combines convolutional feature extraction, spatial downsampling, dense representation learning, and regularization within a unified framework suitable for chest X-ray pneumonia classification. This baseline architecture subsequently serves as the foundation for the Ocotillo Optimization Algorithm-based hyperparameter optimization stage described later in this work.

4.4 Feature Representation

Feature representation constitutes a critical stage in the proposed classification framework because the quality of the extracted representations directly influences the effectiveness of feature selection and subsequent classification performance. In medical image analysis, chest radiographs contain large amounts of spatial information, including texture variations, opacity distributions, and structural lung patterns that may indicate pathological abnormalities. CNN-based representation learning enables these characteristics to be transformed into high-dimensional feature vectors that can later be evaluated and optimized.

In the proposed framework, the CNN was used as a deep representation-learning model to extract discriminative features from chest X-ray images. The convolutional layers automatically learned hierarchical image patterns ranging from low-level edge and texture information to higher-level pathological structures associated with pneumonia. After the final convolutional and pooling operations, the resulting feature maps were flattened into a one-dimensional feature vector suitable for subsequent optimization and classification procedures.

Feature selection was then performed using the Ocotillo Optimization Algorithm to identify the most informative subset of the extracted representations. Redundant or weakly informative features may increase computational complexity and negatively affect classifier generalization. Consequently, selecting a compact subset of discriminative features can improve the efficiency and stability of the classification process while preserving diagnostically relevant information.

Feature selection was applied to deep features extracted from a CNN layer.

The selected deep features were subsequently evaluated using CNN-based and conventional machine-learning classifiers to examine their discriminative capability across multiple classification paradigms. This approach allows the study to investigate whether the optimized feature subsets remain informative beyond a single classifier configuration and supports a broader assessment of feature-selection effectiveness within the proposed framework.

4.5 OcOA-Based Feature Selection

Feature selection was incorporated into the proposed framework to reduce redundant and weakly informative representations extracted from the convolutional neural network. In high-dimensional medical-image feature spaces, not all extracted features contribute equally to the classification process. Some representations may contain irrelevant or overlapping information that increases computational complexity without improving classification capability. Consequently, selecting a compact subset of informative features can improve learning efficiency and strengthen the discriminative quality of the final representation space.

In the present study, the Ocotillo Optimization Algorithm (OcOA) was employed as a binary optimization approach for feature selection. The optimizer searches the feature space iteratively in order to identify subsets that preserve diagnostically relevant information while reducing unnecessary dimensionality. The selected feature subsets were then evaluated using multiple machine-learning classifiers to assess their classification effectiveness.

4.5.1 Binary Encoding

The feature-selection process was formulated as a binary optimization problem in which each candidate solution represents a subset of the extracted deep features. A binary vector was used to encode candidate feature subsets, where the dimensionality of the vector corresponded to the total number of extracted features obtained from the CNN representation layer.

Within the binary representation, each feature was associated with a binary decision value. A value of 1 indicated that the corresponding feature was selected and included in the classification process, whereas a value of 0 indicated that the feature was excluded from the subset. Through this representation, OcOA explored different combinations of selected and non-selected features during the optimization process.

The binary encoding mechanism allows the optimizer to transform the feature-selection problem into a search process over the combinational feature space. This formulation is appropriate for medical-image classification because it enables the algorithm to identify compact feature subsets while maintaining discriminative information associated with pneumonia-related radiographic patterns.

4.5.2 Fitness Function

The fitness evaluation process was designed to assess the quality of each candidate feature subset during optimization. The objective of the fitness function was to balance classification effectiveness with feature reduction. Consequently, the evaluation process considered two principal components: classification error and selected feature ratio.

The classification error component reflects the predictive performance obtained using the selected feature subset. Lower classification error values indicate that the selected features preserve stronger discriminative capability for distinguishing between normal and pneumonia cases. The selected feature ratio represents the proportion of retained features relative to the total feature dimensionality. Smaller feature subsets are generally preferred because they reduce computational complexity and minimize redundancy.

The optimization process therefore aimed to identify feature subsets that simultaneously reduced classification error and limited the number of selected features. This balance is important because selecting an excessively large number of features may preserve redundant information, whereas overly aggressive feature reduction may discard diagnostically relevant representations required for reliable classification.

4.5.3 Feature-Selection Comparison Algorithms

To evaluate the effectiveness of the proposed OcOA-based feature-selection framework, the binary Ocotillo Optimization Algorithm (bOcOA) was compared with several binary metaheuristic optimization algorithms frequently used in feature-selection applications. The comparison algorithms included binary Grey Wolf Optimizer (bGWO), binary Particle Swarm Optimization (bPSO), binary Stochastic Fractal Search (bSFS), binary Whale Optimization Algorithm (bWOA), binary Multiverse Optimization (bMVO), binary Satin Bowerbird Optimizer (bSBO), and binary Genetic Algorithm (bGA).

These optimization algorithms were selected because they represent diverse search strategies with different exploration and exploitation behaviors. Population-based optimization algorithms are commonly employed in feature-selection problems because they can navigate large combinational search spaces without relying on gradient information. Evaluating multiple optimization strategies therefore provides a broader assessment of the effectiveness of bOcOA within the proposed pneumonia-classification framework.

4.5.4 Optimization Parameters

The optimization stage was conducted using a population-based search strategy in which multiple candidate solutions evolved iteratively during the feature-selection process. The optimization configuration included the population size, maximum number of iterations, number of independent runs, binary transfer mechanism, and randomization policy.

The population size was set to 30 candidate solutions in order to maintain a balance between search diversity and computational efficiency. The number of optimization iterations was fixed at 100 iterations to provide sufficient search exploration while avoiding excessive computational overhead. To improve result reliability and reduce sensitivity to stochastic behavior, the optimization process was executed across 30 independent runs.

Because OcOA operates originally within a continuous search space, a binary conversion mechanism was required for feature-selection representation. A sigmoid transfer function was therefore employed to map continuous position values into binary decisions corresponding to feature inclusion or exclusion. The transfer mechanism enabled the optimizer to operate effectively within the binary feature-selection domain.

The optimization process was initialized using randomly generated populations at the beginning of each independent run. The random seed was not fixed in order to evaluate the stability and robustness of the optimization behavior across multiple stochastic executions. This approach allows the reported performance measures to reflect the general optimization tendency rather than the outcome of a single deterministic initialization.

4.6 ML Baseline Models After Feature Selection

After feature selection, the selected feature subsets were evaluated using multiple machine-learning classifiers to assess their discriminative capability across different learning paradigms. Employing several baseline models provides a broader understanding of feature quality because informative representations should remain effective across both deep-learning and conventional machine-learning approaches.

The selected features were evaluated using Convolutional Neural Network (CNN), Neural Network or Multi-Layer Perceptron (MLP), Gradient Boosting, K-Nearest Neighbors (KNN), Naive Bayes, and Random Forest classifiers. These models were selected because they represent distinct classification strategies with different assumptions regarding feature interactions, nonlinear representation learning, and decision-boundary construction.

The CNN baseline used the architecture described previously in Section 4.4. The selected features were supplied to the network for classification learning under the same preprocessing and training conditions adopted throughout the experimental framework. The MLP classifier consisted of fully connected layers with nonlinear activation functions and was trained using backpropagation-based optimization. Gradient Boosting employed an ensemble-learning strategy in which multiple weak learners were combined sequentially to improve predictive performance. KNN performed instance-based classification by assigning labels according to the nearest neighboring feature vectors within the selected feature space. Naive Bayes utilized probabilistic classification based on conditional independence assumptions among features, whereas Random Forest constructed an ensemble of decision trees using randomized feature selection and bootstrap aggregation.

All baseline models were trained using the same feature subsets generated by the feature-selection stage in order to maintain a fair comparison framework. The training procedure involved supervised learning using the training subset of the dataset, while model evaluation was conducted using the testing subset. The validation subset was used during model development and configuration assessment where applicable.

The evaluation process employed multiple classification metrics, including accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F-score. These metrics provide complementary insights into classifier behavior and are particularly important in medical classification tasks where both false-negative and false-positive predictions may have clinical implications. Sensitivity evaluates the ability of the classifier to correctly identify pneumonia cases, whereas specificity measures the correct identification of normal cases. PPV and NPV assess predictive reliability, while F-score provides a balanced evaluation of precision and recall behavior.

4.7 OcOA-Based Hyperparameter Optimization of CNN

After feature selection, the Ocotillo Optimization Algorithm (OcOA) was employed to optimize the hyperparameters of the convolutional neural network. Hyperparameter optimization is an important stage in CNN-based medical image classification because the learning behavior of the model depends strongly on the configuration of its training and architectural parameters. A manually selected configuration may provide acceptable performance, but it does not necessarily represent the most suitable setting for the target dataset. For this reason, OcOA was used to search for an improved CNN configuration by exploring candidate hyperparameter combinations and selecting the configuration that produced the most effective classification behavior.

In the proposed framework, each OcOA candidate solution represented a possible CNN hyperparameter configuration. During the optimization process, the candidate configurations were evaluated according to the classification performance obtained by the corresponding CNN model. The optimizer then updated its candidate solutions iteratively, with the objective of identifying a configuration that improved pneumonia classification from chest X-ray images. This procedure allowed the CNN to be adapted more effectively to the dataset rather than relying only on a fixed manually defined architecture.

4.7.1 Optimized Hyperparameters

The CNN hyperparameters optimized by OcOA included both training-related and architecture-related parameters. The learning rate was optimized within the range 10^{-5} to 10^{-2}, allowing the search process to identify a suitable gradient-update scale for stable convergence. The batch size was selected from the candidate values $\{16, 32, 64\}$, which provided a practical balance between memory consumption and training stability.

The number of convolutional filters was optimized using the candidate values $\{32, 64, 128, 256\}$. These values allowed the optimizer to control the representational capacity of the convolutional layers while avoiding an unnecessarily large model. Kernel size was selected from $\{3\times3, 5\times5\}$, enabling the CNN to learn spatial patterns at different local receptive-field scales. The number of dense units was selected from $\{128, 256, 512\}$, allowing the fully connected component of the network to adapt its representation capacity to the extracted deep features.

The dropout rate was optimized within the range 0.20 to 0.50. This range was used to regulate overfitting by controlling the proportion of deactivated neurons during training. The candidate training optimizers included Adam, RMSprop, and stochastic gradient descent (SGD), each of which provides a different update strategy during gradient-based learning. The number of training epochs was selected from $\{25, 50, 75, 100\}$, allowing OcOA to balance training duration and model convergence.

4.7.2 Search Space

Table 3 presents the CNN hyperparameter search space used by OcOA. The table includes the optimized hyperparameters and their corresponding candidate values or ranges. This search space was designed to include the most influential CNN configuration parameters while maintaining computational feasibility during optimization.

Table 3. CNN hyperparameter search space used by OcOA.

Hyperparameter	Candidate values or range
Learning rate	10^{-5} to 10^{-2}
Batch size	$16, 32, 64$
Number of convolutional filters	$32, 64, 128, 256$
Kernel size	$3\times3, 5\times5$
Dropout rate	0.20 to 0.50
Dense units	$128, 256, 512$
Training optimizer	Adam, RMSprop, SGD
Epochs	$25, 50, 75, 100$

4.7.3 Competing Optimizer-CNN Models

To evaluate the effectiveness of OcOA-based CNN hyperparameter optimization, the proposed OcOA+CNN model was compared with several optimizer-CNN combinations. The competing models included Grey Wolf Optimizer with CNN (GWO+CNN), Genetic Algorithm with CNN (GA+CNN), Particle Swarm Optimization with CNN (PSO+CNN), and Whale Optimization Algorithm with CNN (WOA+CNN). These models were selected because they represent established population-based optimization strategies that are commonly used for hyperparameter search and model configuration tasks.

Each competing optimizer was applied to the CNN hyperparameter optimization problem using the same general search-space design reported in Table 3. This ensured that the comparison focused on the behavior of the optimization algorithms rather than differences in the available hyperparameter choices. After optimization, each optimizer-CNN model was evaluated using the same classification metrics, namely accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F-score.

The comparison between OcOA+CNN and the competing optimizer-CNN models provides an empirical basis for assessing whether OcOA offers an effective search strategy for configuring CNN models in the context of chest X-ray pneumonia classification. By applying the same evaluation protocol across all optimizer-CNN variants, the study maintains a consistent experimental setting for examining the contribution of OcOA to CNN hyperparameter optimization.

5. EXPERIMENTAL SETUP

This section describes the computational environment, training configuration, evaluation criteria, and reporting strategy adopted in the experimental framework. The objective of the experimental setup is to ensure that the proposed methodology can be reproduced and evaluated under a consistent and transparent configuration. Because the study integrates convolutional neural networks, feature-selection procedures, and metaheuristic optimization algorithms within a unified framework, careful specification of the implementation environment and evaluation protocol is necessary to support reliable comparison among the investigated models.

5.1 Software and Hardware

The experimental framework was implemented using Python 3.10 because of its extensive ecosystem for machine learning, deep learning, numerical computation, and medical image analysis. Deep-learning operations associated with the convolutional neural network architecture were implemented using TensorFlow and Keras, while conventional machine-learning models were implemented using the scikit-learn library. Numerical computations and matrix operations were handled using the NumPy library, whereas image preprocessing and image-loading operations were performed using OpenCV and the Python Imaging Library (PIL).

The experiments were executed using a workstation equipped with an Intel Core i7 processor, an NVIDIA RTX-series graphics processing unit (GPU), and 32 GB of RAM.

GPU acceleration was employed to reduce CNN training time and support the iterative optimization procedures associated with OcOA-based hyperparameter tuning. The computational environment therefore provided sufficient resources for deep-learning training, feature extraction, feature-selection optimization, and baseline-model evaluation within a unified experimental framework.

Table 4 summarizes the software and hardware configuration used in this study.

Table 4. Software and hardware environment used in the experiments.

Component	Specification
Programming language	Python 3.10
Deep-learning framework	TensorFlow / Keras
Machine-learning library	scikit-learn
Numerical computation library	NumPy
Image-processing libraries	OpenCV, PIL
CPU	Intel Core i7
GPU	NVIDIA RTX-series GPU
RAM	32 GB

5.2 Training Configuration

The dataset organization followed the original train, test, and validation structure provided by the dataset repository. The training subset was used for CNN learning, feature extraction, and optimization procedures. The validation subset was used during model development and hyperparameter optimization to monitor training behavior and reduce the risk of overfitting. The testing subset was reserved for final evaluation and comparative performance assessment.

The CNN training process was performed using a batch size of 32 and a maximum training duration of 100 epochs. These settings provide a balance between computational efficiency and learning stability during deep-network optimization. Binary cross-entropy was employed as the loss function because the classification problem addressed in this study involves two diagnostic classes, namely *Normal* and *Pneumonia*. The Adam optimization algorithm was used during CNN training because of its adaptive learning-rate mechanism and stable convergence behavior in deep-learning applications.

To improve training stability and reduce unnecessary training iterations, an early-stopping mechanism was incorporated into the learning process. The early-stopping procedure monitored the validation loss during training and terminated the learning process when no improvement was observed for 10 consecutive epochs. This strategy reduces the likelihood of overfitting while preserving the model state associated with the best validation behavior.

Table 5 summarizes the CNN training configuration used throughout the experiments.

Table 5. CNN training configuration used in the experiments.

Parameter	Configuration
Dataset split	Training / Validation / Testing
Batch size	32
Maximum epochs	100
Loss function	Binary cross-entropy
Training optimizer	Adam
Early stopping	Enabled
Patience	10 epochs
Monitoring criterion	Validation loss

5.3 Evaluation Metrics

The proposed framework was evaluated using several classification metrics commonly employed in medical image analysis and binary diagnostic classification. Multiple evaluation measures were adopted because a single metric may not fully capture the predictive behavior of a medical classification system. In particular, medical diagnosis requires balanced evaluation of both positive and negative predictions because false-negative and false-positive outcomes may have important clinical implications.

Accuracy measures the proportion of correctly classified samples relative to the total number of samples and is defined as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives.

Sensitivity, also referred to as the true positive rate (TPR) or recall, evaluates the ability of the classifier to correctly identify pneumonia cases. Sensitivity is defined as:

$$\text{Sensitivity} = \frac{TP}{TP+FN} \tag{2}$$

Specificity, also referred to as the true negative rate (TNR), measures the ability of the classifier to correctly identify normal cases. Specificity is computed as:

$$\text{Specificity} = \frac{TN}{TN+FP} \tag{3}$$

Positive predictive value (PPV), commonly referred to as precision, evaluates the reliability of positive predictions and is defined as:

$$\text{PPV} = \frac{TP}{TP+FP} \tag{4}$$

Negative predictive value (NPV) measures the reliability of negative predictions and is defined as:

$$\text{NPV} = \frac{TN}{TN+FN} \tag{5}$$

The F-score provides a balanced evaluation of precision and recall behavior and is defined as:

$$\text{F-score} = \frac{2\times\text{PPV}\times\text{Sensitivity}}{\text{PPV}+\text{Sensitivity}} \tag{6}$$

These evaluation metrics collectively provide a comprehensive assessment of classifier behavior in pneumonia detection. Accuracy reflects overall predictive correctness, sensitivity evaluates pneumonia-case detection capability, specificity measures normal-case recognition capability, while PPV, NPV, and F-score provide additional insight into predictive reliability and classification balance.

5.4 Reporting Protocol

The experimental findings are organized into three primary result groups in order to maintain a structured and interpretable presentation framework. The first group reports the feature-

Table 6. Feature-selection performance comparison among binary optimization algorithms.

Metric	bOcOA	bGWO	bPSO	bSFS	bWOA	bMVO	bSBO	bGA
Average error	0.523417	0.540617	0.574417	0.584017	0.574217	0.551117	0.582717	0.554217
Average Select size	0.476217	0.676217	0.676217	0.815617	0.839617	0.772717	0.846517	0.618617
Average Fitness	0.586617	0.602817	0.601217	0.624117	0.609017	0.630917	0.640917	0.614217
Best Fitness	0.488417	0.523117	0.581517	0.513817	0.573117	0.556117	0.584017	0.517517
Worst Fitness	0.586917	0.590017	0.649217	0.615417	0.649217	0.674117	0.663717	0.632617
Standard deviation Fitness	0.408917	0.413617	0.413017	0.422917	0.415217	0.463717	0.473917	0.415217

selection results obtained using binary OcOA and the competing binary optimization algorithms. These results evaluate the effectiveness of the optimization algorithms in selecting compact and informative feature subsets.

The second group presents the machine-learning baseline results obtained after feature selection. In this stage, the selected feature subsets are evaluated using CNN and conventional machine-learning classifiers in order to examine the discriminative quality of the optimized representations across multiple classification paradigms.

The third group reports the CNN hyperparameter optimization results obtained using OcOA and the competing optimizer-CNN frameworks. These results assess the contribution of metaheuristic optimization to CNN configuration and final pneumonia-classification behavior.

The Results section focuses on direct reporting of the experimental findings through tables and concise textual observations. Numerical interpretation and broader analytical discussion are reserved primarily for the Discussion section in order to maintain a clear separation between empirical reporting and interpretive analysis. This reporting strategy follows standard academic writing practices commonly adopted in Springer journal manuscripts, where the Results section emphasizes presentation of findings while the Discussion section focuses on interpretation, implications, and contextual analysis.

6. RESULTS

This section presents the experimental findings obtained from the proposed pneumonia-classification framework. The results are reported in three stages. First, the feature-selection performance of binary OcOA is compared with the competing binary optimization algorithms. Second, the selected feature subset is evaluated using CNN and conventional machine-learning baseline models. Third, the optimized CNN models are compared to assess the effect of OcOA-based hyperparameter optimization. The interpretation of these findings is reserved for the Discussion section, while this section focuses on direct empirical reporting.

6.1 Feature-Selection Results

The feature-selection stage evaluated the ability of binary optimization algorithms to identify informative feature subsets from the extracted deep representations. Table 6 reports the feature-selection performance of binary Ocotillo Optimization Algorithm (bOcOA) and the competing binary optimizers, including binary Grey Wolf Optimizer (bGWO), binary Particle Swarm Optimization (bPSO), binary Stochastic Fractal Search (bSFS), binary Whale Optimization Algorithm (bWOA), binary Multiverse Optimization (bMVO), binary Satin Bowerbird Optimizer (bSBO), and binary Genetic Algorithm (bGA).

As shown in Table 6, bOcOA produced the lowest average error among the reported methods. It also produced the smallest average selected feature size, indicating that the selected subset retained fewer features than the subsets generated by the competing binary optimizers. The best fitness value reported for bOcOA was 0.488417. No statistical significance claim is made because statistical testing results are not reported in the present experiment.

The tabular feature-selection results in Table 6 are further supported by the visual summaries in Figs. 2–4. These figures provide complementary views of the same feature-selection stage by showing the metric distribution pattern, inter-algorithm similarity, and average selected feature size across the competing binary optimizers.

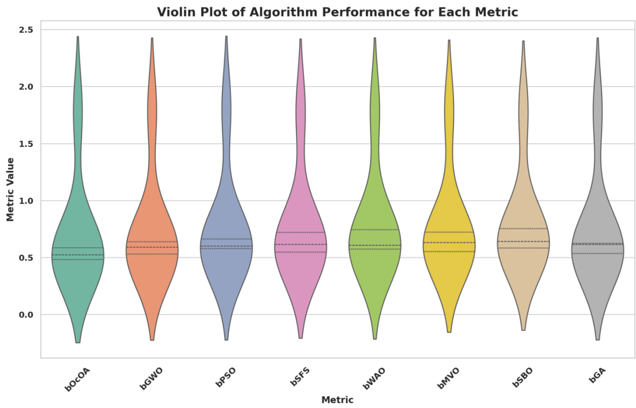

Figure 2. Violin plot of feature-selection algorithm performance across the reported metrics. The figure visualizes the distribution of metric values for bOcOA, bGWO, bPSO, bSFS, bWOA, bMVO, bSBO, and bGA.

Figure 2 presents the distribution of the reported performance metrics for each binary optimization algorithm. The violin plot provides a compact view of how the metric values are spread for each method, allowing the relative behavior of the algorithms to be inspected beyond a single scalar value. In this visualization, bOcOA shows a comparatively lower central tendency for the reported values, which is consistent with its lower average error and smaller selected feature size in Table 6. The plot also shows that the competing algorithms exhibit broadly similar value ranges, although their distributions differ in center and spread across the reported metrics. This visual pattern supports the observation that bOcOA achieved a more compact feature-selection behavior while maintaining competitive metric values across the evaluated criteria.

Figure 3 shows the correlation structure among the binary optimization algorithms based on their reported feature-selection metrics. The heatmap indicates a high degree of similarity among several algorithms, with correlation values close to one for many pairwise comparisons. This suggests

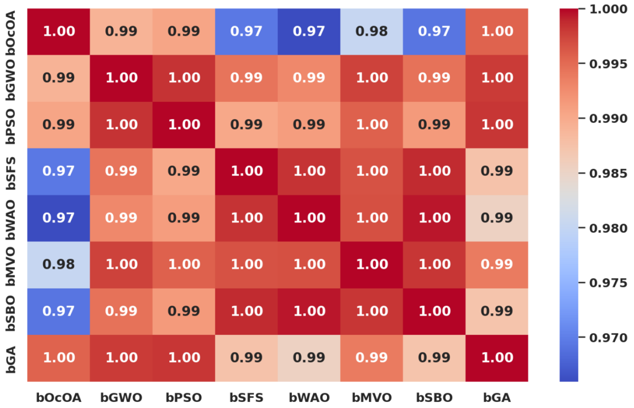

Figure 3. Correlation heatmap of binary optimization algorithms based on the reported feature-selection metrics. The figure summarizes pairwise similarity among bOcOA, bGWO, bPSO, bSFS, bWOA, bMVO, bSBO, and bGA.

that the algorithms follow broadly similar performance trends across the selected metrics, although their absolute values differ. The comparatively lower correlations involving bOcOA and some competing algorithms reflect the distinct feature-selection profile observed for bOcOA, particularly its smaller average selected feature size and lower average error. The heatmap therefore complements Table 6 by showing that the differences among algorithms are not only reflected in individual metric values but also in the overall pattern of metric behavior.

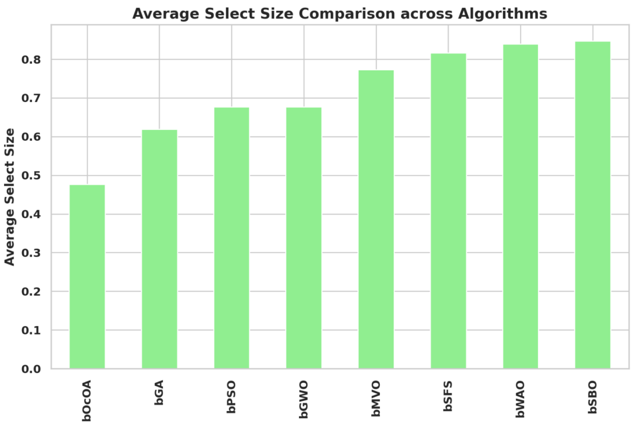

Figure 4. Average selected feature size comparison across the evaluated binary optimization algorithms. The figure highlights the compactness of the feature subset selected by bOcOA relative to the competing binary optimizers.

Figure 4 compares the average selected feature size across the evaluated binary optimization algorithms. This figure is particularly relevant because feature selection aims not only to preserve classification-related information but also to reduce the dimensionality of the selected representation. As shown in the bar chart, bOcOA selected the smallest average feature subset among the reported algorithms, followed by bGA, bPSO, and bGWO. By contrast, bSBO, bWOA, and bSFS retained larger average feature subsets. The visual comparison reinforces the numerical values reported in Table 6, where bOcOA achieved an average selected size of 0.476217. This finding indicates that bOcOA produced the most compact selected representation among the compared methods, although no claim regarding reduced inference time is made because runtime measurements were not reported.

6.2 ML Baseline Results After Feature Selection

After feature selection, the selected feature subset was evaluated using CNN and conventional machine-learning baseline models. Table 7 presents the classification results obtained after feature selection using Convolutional Neural Network (CNN), Neural Network or Multi-Layer Perceptron (MLP), Gradient Boosting, K-Nearest Neighbors (KNN), Naive Bayes, and Random Forest classifiers.

Table 7 presents the classification performance of the baseline models after applying feature selection. CNN achieved the highest accuracy and F-score among the listed baseline models. Random Forest achieved the lowest accuracy among the reported baseline models. The strongest baseline specificity was also obtained by CNN.

The baseline-model results reported in Table 7 are further illustrated through the distributional and clustering visualizations shown in Figs. 5–7. These figures provide complementary perspectives on the behavior of the evaluated machine-learning models after feature selection by examining metric density distributions, cumulative metric behavior, and inter-metric similarity structures.

Figure 5 presents kernel density estimation (KDE) plots for the reported evaluation metrics across the baseline models. The density curves summarize the concentration and spread of the metric values associated with accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F-score. The plots show that the evaluated models generally produced metric distributions concentrated within relatively high-performance regions, particularly for sensitivity and F-score. The specificity distribution exhibits comparatively broader variation, which is consistent with the numerical results reported in Table 7, where specificity values differed more noticeably across the evaluated classifiers. The KDE visualization therefore highlights the relative stability of some metrics while also illustrating the variability associated with normal-case recognition performance among the baseline models.

Figure 6 shows the cumulative distribution behavior of the reported evaluation metrics across the baseline models. The histograms include the mean value, median value, and standard-deviation boundaries for each metric. The visual distributions demonstrate that several metrics, particularly sensitivity and F-score, are concentrated within narrow intervals, indicating relatively consistent behavior across the evaluated models. By contrast, specificity and negative predictive value exhibit comparatively broader distributions, reflecting greater variation among the classifiers in distinguishing normal chest X-ray images from pneumonia cases. The cumulative distributions also show that the CNN baseline is positioned within the upper range of the reported metric distributions, which is consistent with the higher accuracy and F-score values reported previously.

Figure 7 presents a cluster map of the baseline-model performance metrics after feature selection. The cluster visualization groups related metric behaviors according to similarity patterns across the evaluated classifiers. The heatmap and associated dendrogram structure indicate that sensitivity and negative predictive value exhibit relatively similar clustering behavior, while specificity and positive predictive value form distinct grouping patterns. This observation reflects the interaction between pneumonia-case recognition and normal-case

Table 7. Classification performance of ML baselines after feature selection.

Model	Accuracy	Sensitivity (TPR)	Specificity (TNR)
CNN	0.94985	0.955898387096774	0.929016666666667
Neural Network (MLP)	0.92985	0.952672580645161	0.851238888888889
Gradient Boosting	0.91985	0.93976935483871	0.851238888888889
K-Nearest Neighbors	0.91735	0.949446774193548	0.806794444444445
Naive Bayes	0.91485	0.952672580645161	0.784572222222222
Random Forest	0.88985	0.933317741935484	0.740127777777778

Model	PPV	NPV	F-score
CNN	0.952703697749196	0.93987808988764	0.95429847020934
Neural Network (MLP)	0.930804258675079	0.926205421686747	0.9416163476874
Gradient Boosting	0.930401118210863	0.881890229885058	0.935062680577849
K-Nearest Neighbors	0.9186	0.91235	0.933778571428572
Naive Bayes	0.91281439628483	0.923388961038961	0.93233420221169
Random Forest	0.900044704049844	0.848425949367089	0.916391204437401

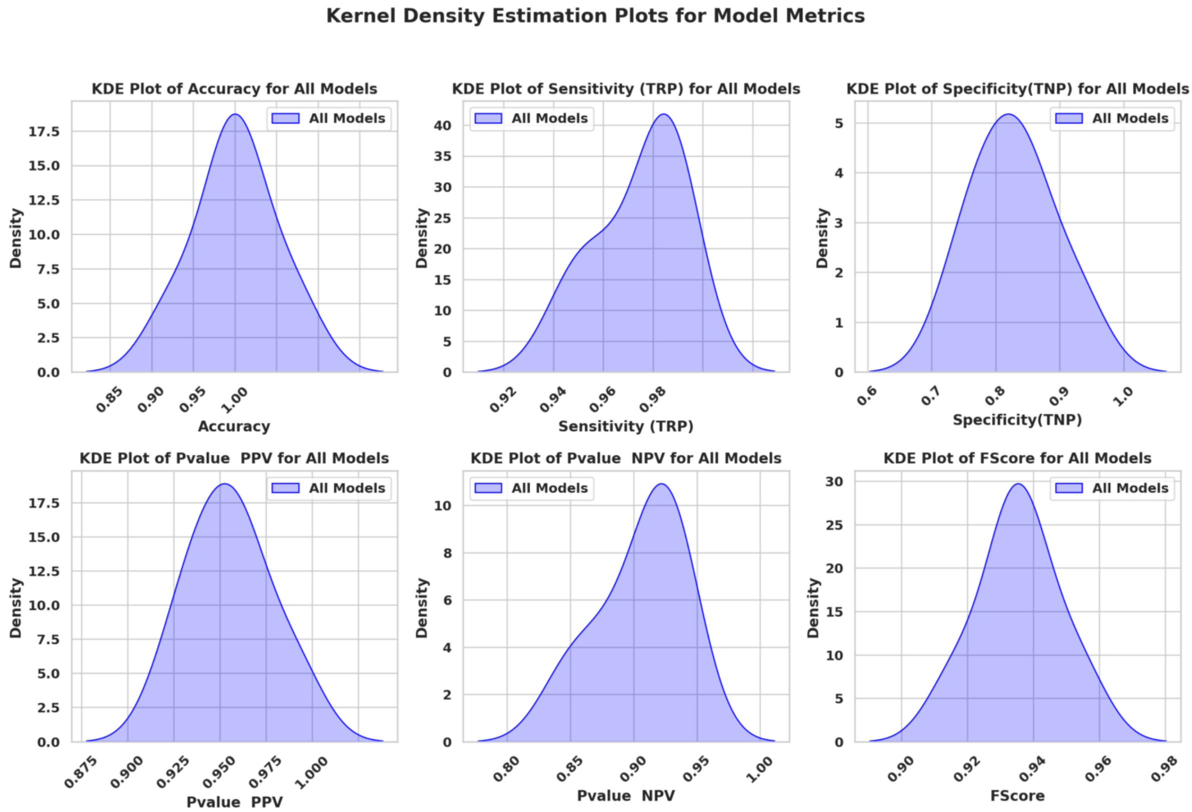

Figure 5. Kernel density estimation plots for the evaluation metrics obtained by the baseline models after feature selection. The figure summarizes the metric distributions for accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F-score across the evaluated classifiers.

recognition across the investigated models. The cluster map also illustrates that the evaluation metrics are not completely independent; instead, several metrics exhibit correlated behavior patterns across the baseline classifiers. Such visualization complements the tabular results by highlighting structural relationships among the reported performance measures.

6.3 Hyperparameter Optimization Results

The final stage evaluated CNN models optimized using different metaheuristic algorithms. Table 8 reports the classification performance of the optimized CNN models, including OcOA+CNN, GWO+CNN, GA+CNN, PSO+CNN, and WOA+CNN.

As reported in Table 8, OcOA+CNN achieved the highest accuracy, sensitivity, specificity, PPV, NPV, and F-score among the reported optimizer-CNN models. The best reported accuracy was 0.989939, obtained by OcOA+CNN. The best reported F-score was 0.99438747020934, also obtained by OcOA+CNN.

The optimizer-CNN results reported in Table 8 are further examined through the visual summaries in Figs. 8–10. These figures provide complementary views of the optimized model behavior by illustrating metric dispersion, model-wise metric trends, and the cumulative contribution of the reported evaluation measures across the competing optimizer-CNN frameworks.

Figure 8 presents histograms with mean and standard-deviation indicators for the evaluation metrics obtained by the optimized CNN models. The figure summarizes the distri-

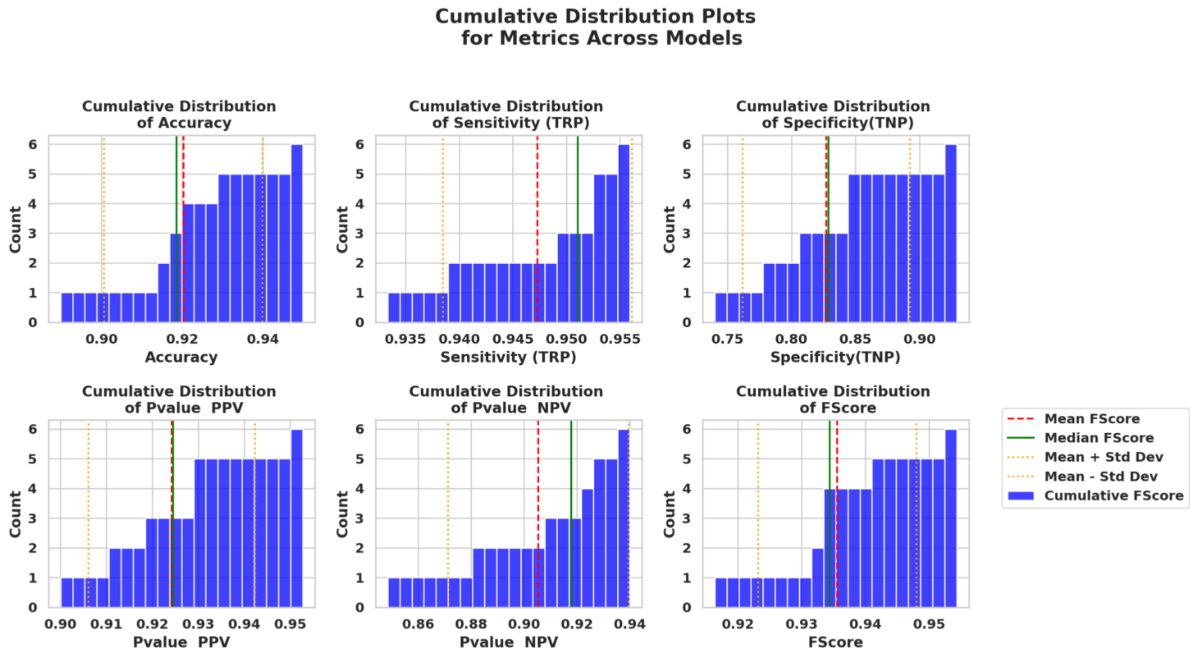

Figure 6. Cumulative distribution plots for the evaluation metrics obtained by the baseline models after feature selection. The figure includes mean, median, and standard-deviation boundaries for each reported metric.

Table 8. Classification performance of optimizer-CNN models.

Model	Accuracy	Sensitivity (TPR)	Specificity (TNR)
OcOA+CNN	0.989939	0.995987387096774	0.969105666666667
GWO+CNN	0.969939	0.992761580645161	0.891327888888889
GA+CNN	0.959939	0.97985835483871	0.891327888888889
PSO+CNN	0.957439	0.989535774193548	0.846883444444445
WOA+CNN	0.954939	0.992761580645161	0.824661222222222

Model	PPV	NPV	F-score
OcOA+CNN	0.992792697749196	0.979967089887641	0.99438747020934
GWO+CNN	0.970893258675079	0.966294421686747	0.9817053476874
GA+CNN	0.970490118210863	0.921979229885058	0.975151680577849
PSO+CNN	0.958689	0.952439	0.973867571428572
WOA+CNN	0.95290339628483	0.963477961038961	0.97242320221169

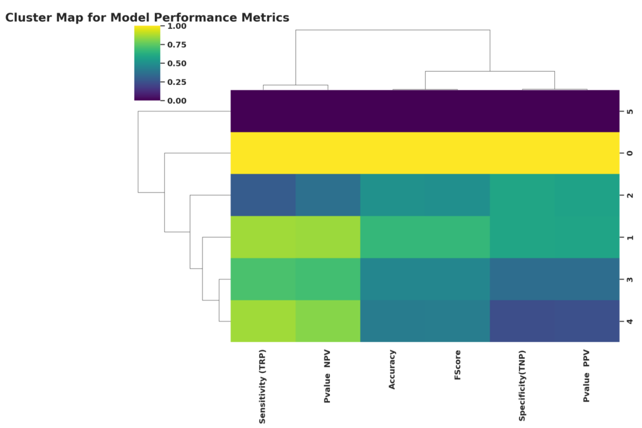

Figure 7. Cluster map of the evaluation metrics obtained by the baseline models after feature selection. The figure illustrates similarity relationships among the reported metrics across the evaluated classifiers.

bution of accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and F-score across OcOA+CNN, GWO+CNN, GA+CNN, PSO+CNN, and WOA+CNN. The histograms show that sensitivity values are concentrated within a narrow high-performance interval, whereas specificity displays wider variation among the optimized models. This pattern is consistent with Table 8, where the optimized models generally achieved high pneumonia-case detection capability, while normal-case recognition differed more substantially across optimizers.

Figure 9 shows the metric trends across the optimized CNN models using line plots with confidence-interval-style bands. The visualization allows direct inspection of how each metric changes from one optimizer-CNN framework to another. OcOA+CNN appears at the upper end of the plotted metric trajectories for the reported performance measures, which agrees with the values presented in Table 8. The line plot also highlights that sensitivity remains consistently high across the optimized models, whereas specificity declines more clearly for several competing optimizers. This difference indicates that the main variability among the optimized CNN models is associated more with normal-case recognition than pneumonia-case detection.

Figure 10 presents a stacked bar chart of the reported evaluation metrics for each optimized CNN model. This visualization provides an aggregate view of the metric profile associ-

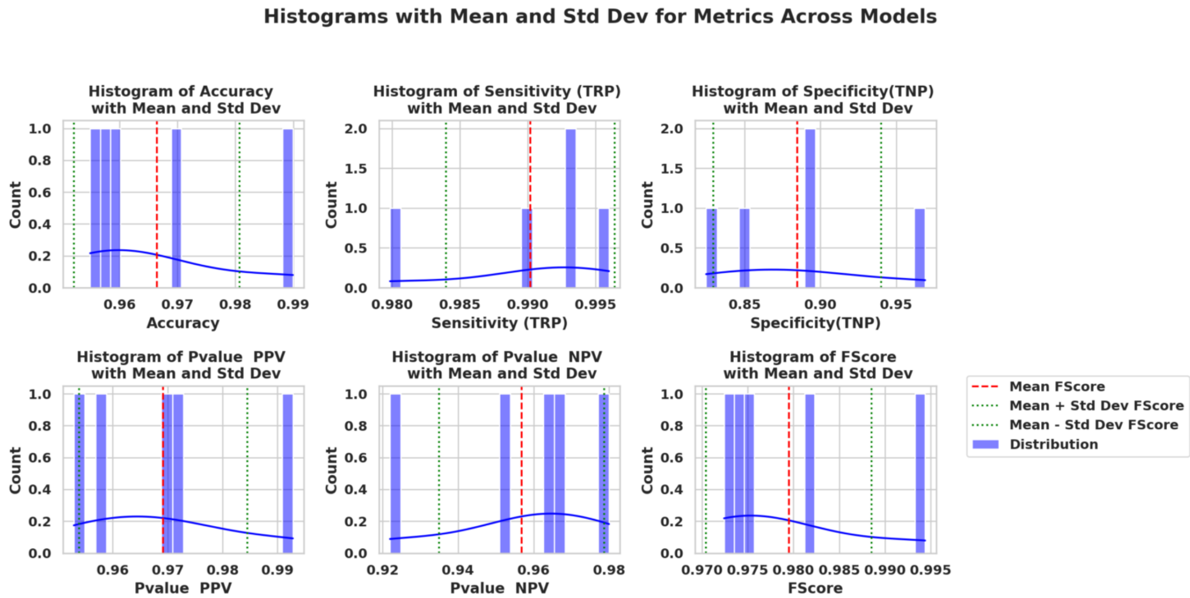

Figure 8. Histograms with mean and standard-deviation indicators for the evaluation metrics obtained by the optimized CNN models. The figure summarizes the distributions of accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F-score across the optimizer-CNN frameworks.

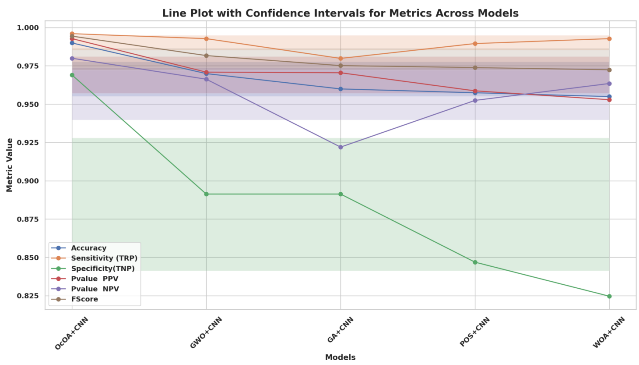

Figure 9. Line plot with confidence-interval-style bands for the evaluation metrics across the optimized CNN models. The figure illustrates model-wise trends for accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F-score.

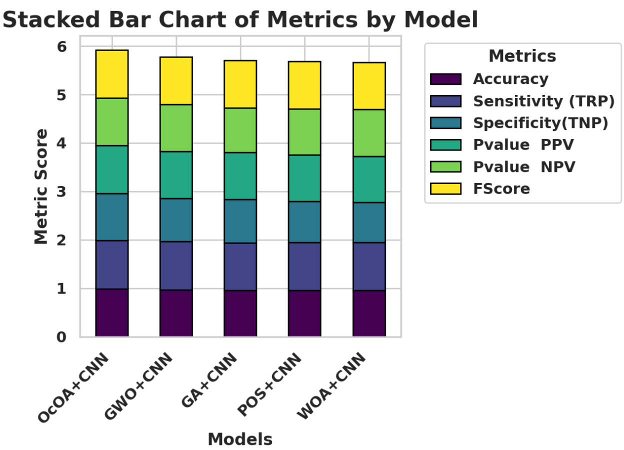

Figure 10. Stacked bar chart of the evaluation metrics for the optimized CNN models. The figure provides an aggregate comparison of accuracy, sensitivity, specificity, positive predictive value, negative predictive value, and F-score across OcOA+CNN, GWO+CNN, GA+CNN, PSO+CNN, and WOA+CNN.

ated with each optimizer-CNN framework. The OcOA+CNN bar has the largest combined metric height among the compared models, reflecting its stronger overall performance profile in Table 8. The figure also shows that the difference among optimizer-CNN models is not driven by a single metric alone; rather, the relative advantage of OcOA+CNN is supported by consistently high values across the reported evaluation measures. This visual summary reinforces the tabular finding that OcOA-based CNN hyperparameter optimization produced the strongest optimized classification behavior within the evaluated experimental setting.

6.4 Summary of Main Empirical Findings

The empirical findings reported in Tables 6–8 show a consistent pattern across the three experimental stages. In the feature-selection comparison, bOcOA produced the smallest selected feature ratio and the lowest average error among the reported binary optimization algorithms. In the baseline classification stage, CNN was the strongest model after feature selection according to the reported accuracy and F-score. In the hyperparameter optimization stage, OcOA+CNN was the strongest optimized model across all reported evaluation metrics.

The optimized OcOA+CNN model also improved over the CNN baseline after feature selection in terms of accuracy and F-score. These findings provide the empirical basis for the interpretive discussion presented in the following section.

7. DISCUSSION

This section discusses the implications of the experimental findings obtained from the proposed OcOA-CNN framework for pneumonia classification from chest X-ray images. The discussion focuses on the behavior of the feature-selection process, the classification performance of the baseline models, the contribution of OcOA-based hyperparameter optimization, and the broader practical and methodological implications of the study. The interpretation is restricted to the experimental evidence reported in the Results section and does not extend to unsupported claims beyond the evaluated dataset and experimental configuration.

7.1 Interpretation of Feature-Selection Results

The feature-selection results indicate that binary Ocotillo Optimization Algorithm (bOcOA) produced the smallest average selected feature size among the investigated binary optimization algorithms. This observation is important because high-dimensional feature spaces extracted from convolutional neural networks frequently contain redundant or weakly informative representations. Retaining unnecessarily large feature subsets may complicate the learning process and increase the dimensionality of the classification space without contributing additional discriminative value.

The ability of bOcOA to select a comparatively smaller feature subset suggests that the optimization process was capable of identifying compact representations while preserving classification-related information. A reduced feature subset may improve the interpretability and manageability of the learned representation space by limiting the inclusion of overlapping or less informative features. In addition, compact feature subsets may reduce the computational burden associated with model training and optimization because fewer feature dimensions are processed during classification. However, no explicit inference-time measurements or runtime-complexity analyses were conducted in the present study; therefore, claims regarding reductions in inference latency or computational speed are intentionally avoided.

The reported average error values also indicate that the feature subsets generated by bOcOA preserved discriminative information effectively relative to the competing binary optimization algorithms. Because feature selection was performed on deep features extracted from the CNN representation layer, the optimization process operated within a high-dimensional and nonlinear representation space. The observed behavior of bOcOA therefore suggests that the optimizer was capable of balancing feature reduction with classification-oriented feature preservation.

Another notable observation concerns the relative behavior of the competing binary optimization algorithms. Algorithms such as bSFS, bWOA, and bSBO produced larger average selected feature sizes than bOcOA, which may indicate a tendency toward retaining broader feature subsets during the search process. Similarly, the reported fitness values demonstrate variability across the investigated algorithms, reflecting differences in their exploration and exploitation behaviors during optimization. Since the present work does not include convergence analysis or statistical significance testing, the interpretation remains limited to the empirical patterns observed in the reported results.

7.2 Interpretation of ML Baseline Results

The machine-learning baseline evaluation demonstrates that CNN achieved the strongest classification performance among the investigated baseline models after feature selection. This behavior is consistent with the characteristics of convolutional neural networks in medical image analysis because CNNs are specifically designed to preserve and learn spatial relationships within image data. Even when feature selection is incorporated, CNN-based architectures remain effective in capturing hierarchical radiographic patterns associated with pulmonary abnormalities.

Traditional machine-learning classifiers such as Random Forest, K-Nearest Neighbors (KNN), Gradient Boosting, and Naive Bayes generally operate on vectorized feature representations and do not inherently preserve spatial image structure in the same manner as CNNs. Consequently, these classifiers may become more dependent on the discriminative quality of the extracted feature vectors and less capable of capturing subtle local spatial dependencies associated with pneumonia-related opacification patterns.

The Multi-Layer Perceptron (MLP) produced comparatively strong performance among the conventional machine-learning models, which may be attributed to its nonlinear representation-learning capability through fully connected neural processing. Nevertheless, its performance remained below that of CNN, suggesting that the convolutional operations and spatial feature extraction mechanisms within CNN contributed additional representational advantages for chest X-ray classification.

Gradient Boosting also demonstrated competitive behavior relative to the other classical models. Ensemble-learning approaches often perform effectively in structured classification tasks because they combine multiple weak learners iteratively. However, the obtained metrics indicate that the Gradient Boosting classifier remained less effective than CNN-based learning in the present experimental setting.

KNN produced relatively strong sensitivity values, suggesting that the selected feature subsets preserved meaningful neighborhood relationships among the samples. However, the lower specificity values indicate that the classifier was comparatively less effective in distinguishing normal chest X-ray images from pneumonia cases. Because KNN classification depends heavily on local distance relationships in the feature space, the classifier may be more sensitive to overlapping class distributions within high-dimensional medical-image representations.

Naive Bayes produced competitive sensitivity performance but lower specificity compared with CNN and MLP. Since Naive Bayes assumes conditional independence among features, its classification capability may become constrained when complex nonlinear dependencies exist between deep radiographic representations. Chest X-ray images often contain interdependent spatial patterns, and such dependencies may not be fully captured through probabilistic independence assumptions.

Random Forest achieved the lowest reported accuracy among the investigated baseline models. Although Random Forest is generally robust in many classification applications, the observed behavior in the present study suggests that the selected deep feature representations may be more effectively exploited through neural-network-based architectures than through tree-based ensemble classification.

7.3 Interpretation of Hyperparameter Optimization Results

The hyperparameter optimization results indicate that the OcOA-optimized CNN model achieved improved classification performance relative to the non-optimized CNN baseline. Specifically, the classification accuracy increased from 0.94985 to 0.989939, while the F-score increased from 0.95429847020934 to 0.99438747020934. Similarly, specificity increased from 0.929016666666667 to 0.969105666666667 following OcOA-based optimization.

These improvements suggest that the optimization process contributed positively to the CNN configuration by identifying more suitable hyperparameter combinations for the pneumonia-classification task. Hyperparameters such as learning rate, batch size, number of filters, and dropout configuration strongly influence learning dynamics, convergence stability, and representation quality within deep-learning models. Consequently, automated optimization strategies may improve the ability of the CNN to adapt to the statistical structure of chest X-ray images.

The increase in sensitivity is particularly relevant in pneumonia detection because false-negative predictions correspond to pneumonia cases that remain undetected by the classification system. In clinical screening scenarios, missed pneumonia cases may delay treatment or additional medical evaluation. Therefore, maintaining high sensitivity is important when designing automated diagnostic-support systems.

Specificity is also clinically relevant because false-positive predictions may increase unnecessary follow-up examinations or additional diagnostic procedures for healthy individuals. The observed improvement in specificity therefore indicates that the optimized CNN model was more effective in distinguishing between normal chest X-ray images and pneumonia-related abnormalities within the studied dataset.

The improvement in F-score further indicates that the optimization process improved the balance between precision and recall behavior. Since the F-score incorporates both predictive precision and sensitivity, higher values suggest more balanced classification behavior across the two diagnostic categories.

7.4 Comparison with Competing Optimizer-CNN Models

The optimizer-CNN comparison demonstrates that OcOA+CNN achieved higher reported performance metrics than GWO+CNN, GA+CNN, PSO+CNN, and WOA+CNN within the evaluated experimental setting. According to the reported results, OcOA+CNN produced the highest accuracy, sensitivity, specificity, PPV, NPV, and F-score among the investigated optimizer-CNN frameworks.

GWO+CNN produced the second-highest reported accuracy and F-score values among the compared models. This behavior suggests that Grey Wolf Optimizer-based hyperparameter search also provided effective CNN configurations for the chest X-ray classification task. Nevertheless, the metrics reported for OcOA+CNN remained higher across all reported evaluation measures.

GA+CNN and PSO+CNN also demonstrated strong classification capability, although their specificity and F-score values remained below those obtained by OcOA+CNN. These observations suggest differences in the optimization behavior of the investigated metaheuristic algorithms when exploring the CNN hyperparameter search space.

WOA+CNN achieved comparatively high sensitivity values but lower specificity relative to OcOA+CNN. This pattern indicates that the model retained strong pneumonia-detection capability but produced more false-positive predictions compared with the OcOA-optimized model.

It is important to emphasize that the present comparison is limited strictly to the optimizer-CNN frameworks evaluated in this study. No direct comparison with external state-of-the-art methods is performed because external benchmark results, implementation details, and evaluation protocols were not incorporated into the current experimental framework.

7.5 Practical Implications

The proposed OcOA-CNN framework may provide practical value as a computer-aided decision-support approach for chest X-ray screening. Automated pneumonia-detection systems may assist clinicians by providing rapid preliminary assessments and supporting diagnostic workflows in environments where radiological expertise is limited or imaging workloads are high.

The integration of feature selection and hyperparameter optimization within a unified framework may also contribute to improving automated classification quality on the studied dataset. By combining CNN-based representation learning with optimization-driven feature and parameter selection, the framework attempts to address both representation quality and model configuration simultaneously.

Despite these observations, clinical deployment of the proposed framework would require substantially broader validation. External evaluation on independent datasets collected from different medical centers and imaging devices would be necessary to assess generalization capability. In addition, calibration analysis and prospective clinical evaluation would be required before considering deployment in real-world diagnostic settings.

7.6 Threats to Validity

Several factors may influence the validity and generalizability of the findings reported in this study.

7.6.1 Internal Validity

The optimization and training procedures employed in this study involve stochastic initialization processes. Consequently, the obtained results may be partially sensitive to random initialization conditions, optimization trajectories, and data-ordering effects during training. Although multiple optimization runs were conducted during feature selection, random seeds were not fixed throughout the experiments. Future studies should explicitly report seed-control strategies and investigate result stability across multiple controlled executions.

7.6.2 External Validity

The dataset used in this study consists of pediatric chest X-ray images collected from a specific medical institution. As a result, the reported findings may not necessarily generalize to adult populations, alternative imaging protocols, or radiographs collected from other hospitals and imaging devices. Differences in patient demographics, acquisition conditions, and radiographic quality may affect the behavior of machine-learning models trained on a single-source dataset. External validation using independent multi-institutional datasets would therefore be necessary to evaluate broader generalization capability.

7.6.3 Construct Validity

The experimental formulation adopted in this study focuses on binary classification between normal and pneumonia cases. Although the dataset documentation discusses bacterial and viral pneumonia patterns, these subtypes were not treated as separate experimental classes in the present framework. Consequently, the reported results should not be interpreted as evidence of subtype-level pneumonia classification capability.

7.6.4 Statistical Validity

The present study reports empirical classification metrics without performing formal statistical significance testing. Consequently, the observed performance differences among the investigated optimization algorithms and classifier configurations should not be interpreted as statistically significant improvements. Future work should incorporate statistical validation procedures such as Wilcoxon signed-rank testing, Friedman testing, or paired t-tests in order to support stronger comparative conclusions.

8. LIMITATIONS

Several limitations should be acknowledged when interpreting the findings of this study. First, the dataset contains only two experimental classes, namely *Normal* and *Pneumonia*. Although this binary formulation is useful for evaluating pneumonia detection, it does not capture the broader diversity of pulmonary diseases encountered in clinical practice.

Second, the dataset consists of pediatric chest X-ray images collected from a specific medical source. Consequently, the findings reported in this study may not generalize directly to adult populations or to radiographs obtained from different imaging environments and healthcare institutions.

Third, external validation using independent datasets was not included in the present experimental framework. The reported findings are therefore limited to the evaluated dataset and should not be interpreted as evidence of generalized clinical deployment capability.

Fourth, runtime analysis, convergence-time evaluation, and inference-latency measurements were not included in the current study. Although feature selection may reduce representation dimensionality, the present work does not provide direct evidence regarding computational acceleration or real-time deployment efficiency.

Fifth, statistical significance testing was not performed. Therefore, the observed differences between the investigated optimization algorithms and classifier configurations should be interpreted as empirical observations rather than statistically validated superiority claims.

Finally, reproducibility depends strongly on detailed reporting of architectural and optimization settings. Although the present work provides the CNN configuration and OcOA search-space design used in the experiments, future studies may further strengthen reproducibility through open-source implementation release and controlled seed reporting.

9. CONCLUSION

This paper evaluated an OcOA-CNN framework for pneumonia detection from chest X-ray images using deep feature selection and CNN hyperparameter optimization. The proposed framework integrated convolutional neural network-based representation learning with the Ocotillo Optimization Algorithm in two principal stages: feature selection and hyperparameter optimization.

In the feature-selection stage, binary Ocotillo Optimization Algorithm (bOcOA) achieved the lowest reported average error and the smallest average selected feature size among the investigated binary optimization algorithms. These observations suggest that bOcOA was capable of generating compact feature subsets while preserving classification-related information extracted from chest X-ray images.

The selected feature subsets were subsequently evaluated using CNN and conventional machine-learning baseline models. Among the investigated baseline classifiers, CNN achieved the strongest reported classification performance according to the evaluated metrics. This behavior is consistent with the ability of convolutional neural networks to preserve spatial image information and learn hierarchical radiographic representations relevant to pneumonia classification.

In the hyperparameter optimization stage, OcOA+CNN achieved the highest reported classification performance among the investigated optimizer-CNN frameworks. The optimized model achieved an accuracy of 0.989939, sensitivity of 0.995987387096774, specificity of 0.969105666666667, and F-score of 0.99438747020934. These findings indicate that OcOA-based optimization contributed positively to CNN configuration within the evaluated experimental setting.

Future research directions may include external validation using independent multi-institutional datasets, incorporation of statistical significance testing, runtime and convergence analysis, and evaluation of multi-class pneumonia classification scenarios if subtype annotations become available. Additional investigation into explainable artificial intelligence techniques and clinically interpretable feature analysis may also improve the practical applicability of optimization-guided deep-learning frameworks for medical image analysis.

REFERENCES

[1] World Health Organization, “Pneumonia in children,” WHO fact sheet, 2022.

[2] T. Cherian, E. K. Mulholland, J. B. Carlin, H. Ostensen, R. Amin, M. de Campo, D. Greenberg, R. Lagos, M. Lucero, S. A. Madhi, K. L. O’Brien, S. Obaro, and M. C. Steinhoff, “Standardized interpretation of paediatric chest radiographs for the diagnosis of pneumonia in epidemiological studies,” *Bulletin of the World Health Organization*, vol. 83, no. 5, pp. 353–359, 2005.

[3] N. Fancourt, M. Deloria Knoll, B. Barger-Kamate, M. de Campo, J. de Campo, M. B. Diallo, J. Figueroa, L. L. Hammitt, M. M. Higdon, J. K. Igbo, H. L. Johnson, O. S. Levine, N. Lufesi, W. B. MacLeod, S. A. Madhi, N. Mahomed, D. P. Moore, L. Mwananyanda, K. L. O’Brien, J. A. G. Scott, D. M. Thea, S. L. Zeger, and D. R. Feikin, “Standardized interpretation of chest radiographs in cases of pediatric pneumonia from the perch study,” *Clinical Infectious Diseases*, vol. 64, no. suppl_3, pp. S253–S261, 2017.

[4] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, “Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3462–3471, 2017.

[5] P. Rajpurkar, J. Irvin, K. Zhu, B. Yang, H. Mehta, T. Duan, D. Ding, A. Bagul, C. P. Langlotz, K. Shpanskaya, M. P. Lungren, and A. Y. Ng, “Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning,” *arXiv preprint arXiv:1711.05225*, 2017.

[6] P. Mooney, “Chest x-ray images (pneumonia),” Kaggle dataset, 2018.

[7] D. S. Kermany, M. Goldbaum, W. Cai, C. C. S. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. K. Prasadha, J. Pei, M. Y. L. Ting, J. Zhu, C. Li, S. Hewett, J. Dong, I. Ziyar, A. Shi, R. Zhang, L. Zheng,

R. Hou, W. Shi, X. Fu, Y. Duan, V. A. N. Huu, C. Wen, E. D. Zhang, C. L. Zhang, O. Li, X. Wang, M. A. Singer, X. Sun, J. Xu, A. Tafreshi, M. A. Lewis, H. Xia, and K. Zhang, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131.e9, 2018.

[8] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical Image Analysis*, vol. 42, pp. 60–88, 2017.

[9] H.-C. Shin, H. R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, and R. M. Summers, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[10] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Computers & Electrical Engineering*, vol. 40, no. 1, pp. 16–28, 2014.

[11] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *Journal of Machine Learning Research*, vol. 13, pp. 281–305, 2012.

[12] L. Yang and A. Shami, "On hyperparameter optimization of machine learning algorithms: Theory and practice," *Neurocomputing*, vol. 415, pp. 295–316, 2020.

[13] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey wolf optimizer," *Advances in Engineering Software*, vol. 69, pp. 46–61, 2014.

[14] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Networks*, vol. 4, pp. 1942–1948, 1995.

[15] S. Mirjalili and A. Lewis, "The whale optimization algorithm," *Advances in Engineering Software*, vol. 95, pp. 51–67, 2016.

[16] S. Mirjalili, S. M. Mirjalili, and A. Hatamlou, "Multi-verse optimizer: A nature-inspired algorithm for global optimization," *Neural Computing and Applications*, vol. 27, no. 2, pp. 495–513, 2016.

[17] S. H. Samareh Moosavi and V. Khatibi Bardsiri, "Satin bowerbird optimizer: A new optimization algorithm to optimize anfis for software development effort estimation," *Engineering Applications of Artificial Intelligence*, vol. 60, pp. 1–15, 2017.

[18] J. H. Holland, *Adaptation in Natural and Artificial Systems*. Ann Arbor, MI: University of Michigan Press, 1975.

[19] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*. Reading, MA: Addison-Wesley, 1989.

[20] E.-S. M. El-Kenawy, F. H. Rizk, A. M. Zaki, M. E. Mohamed, A. Ibrahim, A. A. Abdelhamid, N. Khodadadi, E. M. Almetwally, and M. M. Eid, "Ocotillo optimization algorithm (ocoa): A desert-inspired metaheuristic for adaptive optimization," *Journal of Artificial Intelligence and Metaheuristics*, no. 1, pp. 39–59, 2025.

[21] J. Kennedy and R. C. Eberhart, "A discrete binary version of the particle swarm algorithm," in *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, vol. 5, pp. 4104–4108, 1997.

[22] H. Salimi, "Stochastic fractal search: A powerful metaheuristic algorithm," *Knowledge-Based Systems*, vol. 75, pp. 1–18, 2015.