Journal of King Saud University - Computer and Information Sciences 36 (2024) 102130

HOSTED BY

Contents lists available at ScienceDirect

Journal of King Saud University - Computer and Information Sciences

journal homepage: www.sciencedirect.com

Full Length Article

An optimized fusion of deep learning models for kidney stone detection from CT images

Sohaib Asif [a,*], Xiaolong Zheng [b,**], Yusen Zhu [c]

[a] School of Computer Science and Engineering, Central South University, Changsha, China
[b] Xi'an Research Institute of High Tech, Xi'an, China
[c] School of Mathematics, Hunan University, Changsha, China

ARTICLE INFO

Keywords:
Kidney stone detection
Deep neural networks
Stack ensemble
Particle swarm optimization
Biomedical image classification

ABSTRACT

Accurate diagnosis of kidney disease is crucial, as it is a significant health concern that demands precise identification for effective and appropriate treatment. Deep learning methods are increasingly recognized as valuable tools for disease diagnosis in the biomedical field. However, current models utilizing deep networks often encounter challenges of overfitting and low accuracy, necessitating further refinement for optimal performance. To overcome these challenges, this paper proposes the introduction of two ensemble models designed for kidney stone detection in CT images. The first model, called StackedEnsembleNet, is a two-level deep stack ensemble model that effectively integrates the predictions from four base models: InceptionV3, InceptionResNetV2, MobileNet, and Xception. By leveraging the collective knowledge of these models, StackedEnsembleNet improves the accuracy and reliability of kidney stone detection. The second model PSOWeightedAvgNet, leverages the Particle Swarm Optimization (PSO) algorithm to determine the optimal weights for the weighted average ensemble. Through PSO, this ensemble approach assigns optimized weights to each model during the ensembling process, effectively enhancing the performance by optimizing the combination of their predictions. Experimental results conducted on a large dataset of 1799 CT images demonstrate that both StackedEnsembleNet and PSOWeightedAvgNet outperform the individual base models, achieving high accuracy rates in kidney stone detection. Furthermore, additional experiments on an unseen dataset validate the models' ability to generalize. The comparison with previous methods confirms the superior performance of the proposed ensemble models. The paper also presents Grad-CAM visualizations and error case analysis to provide insights into the decision-making processes of the models. By overcoming the limitations of existing deep learning models, StackedEnsembleNet and PSOWeightedAvgNet offer a promising approach for accurate kidney stone detection, contributing to improved diagnosis and treatment outcomes in the field of nephrology.

1. Introduction

Kidney diseases pose a significant threat to human health, leading to the loss of kidney function and potential mortality if left untreated (Vupputuri et al., 2004; Alelign and Petros, 2018). Timely identification plays a vital role in averting the advancement of kidney disorders, including conditions like hydronephrosis, cysts, and stones (Edvardsson et al., 2013). Unfortunately, the prevalence of kidney-related disorders is on the rise globally, while the availability of nephrologist's remains limited in many countries (Sorokin et al., 2017). This situation creates a pressing need for automated medical applications to assist in the timely and accurate diagnosis of kidney diseases. The role of medical imaging is crucial in both detecting and diagnosing kidney abnormalities. Different modalities are employed to analyze patients and identify renal conditions (Brisbane et al., 2016). Ultrasound imaging, although safe and radiation-free, is often hindered by limitations in image quality, particularly in cases involving deep-seated tissues and fat (Polat et al., 2017). On the other hand, CT imaging provides high-resolution, contrast-enhanced images, allowing for the precise identification of kidney stones and their characteristics (Shlipak et al., 2005; Asif et al., 2023). CT has the advantage of detecting even small stones, as small as 1 mm, which may be missed by ultrasound. Hence, advancements in

* Corresponding authors at: School of Computer Science and Engineering, Central South University, Changsha, China.
** Corresponding authors at: Xi'an Research Institute of High Tech, Xi'an, China.
E-mail addresses: punjabians1592@gmail.com (S. Asif), Xiaolong.zheng.7712@outlook.com (X. Zheng), zhu_yusen@163.com (Y. Zhu).

https://doi.org/10.1016/j.jksuci.2024.102130
Received 1 May 2024; Received in revised form 7 June 2024; Accepted 10 July 2024
Available online 18 July 2024

imaging techniques, coupled with the increasing utilization of CT, have contributed to the detection of previously undiagnosed small kidney stones (Baygin et al., 2022).

In addition to kidney stones, other kidney abnormalities such as cysts and tumors also significantly impact kidney function and overall health. The formation and enlargement of fluid-filled cysts in the kidneys can compress the nephrons and impair kidney function (Gunasekara et al., 2022). Moreover, renal cell carcinoma (RCC), commonly referred to as kidney tumor, is a rapidly growing malignancy that contributes to a considerable number of cancer-related deaths worldwide (Sung et al., 2021). Timely detection and appropriate management of these abnormalities are essential for preserving kidney function and preventing the development of chronic kidney diseases. Given the increasing burden of kidney diseases and the limited availability of nephrologists, computer-aided medical systems have emerged as a promising approach to assist in the early detection and diagnosis of kidney abnormalities (Asif et al., 2022). Artificial intelligence (AI)-based systems, leveraging machine learning algorithms, have the potential to alleviate the workload on clinicians, minimize human errors, and enhance diagnostic accuracy. By incorporating these systems into routine clinical practice, healthcare professionals can obtain accurate and robust results, ultimately improving patient outcomes. Deep learning (DL) has proven to be a potent and versatile tool, finding applications in diverse fields, including medical image analysis and diagnosis. Its ability to automatically learn intricate patterns and representations from complex data has revolutionized the field (Shen et al., 2017). Within DL, ensemble models hold particular promise. Ensemble models combine multiple individual models, leveraging their collective wisdom to improve predictive accuracy and robustness. By harnessing the diverse strengths of different models, ensemble techniques can overcome the limitations of individual models, reduce overfitting, enhance generalization, and provide more reliable and accurate predictions (Yang et al., 2022). Ensemble models (Asif et al., 2023) have shown remarkable potential in various medical applications, including kidney stone detection, by significantly enhancing diagnostic performance and supporting healthcare professionals in making informed decisions.

Over the past few years, DL techniques have gained significant momentum in the field of urology (Asif et al., 2023; Asif et al., 2023), particularly for automating the detection of kidney stones. Despite the notable advancements of DL techniques in various domains, the utilization of these methods in the diagnosis of kidney diseases has been relatively scarce and lacking in extensive research. In (Wu and Yi, 2020), the authors proposed a feature fusion-based model to classify kidney stones. They trained and evaluated the model on a dataset of 3733 images, achieving an impressive accuracy of 94.67 %. The authors (Parakh et al., 2019) conducted a study using a Inception-v3 network with Softmax classification for kidney stone detection on a dataset comprising 535 CT images. Their approach yielded a high accuracy of 95 %. Yildirim et al. (Yildirim et al., 2021) employed the XResNet-50 model for the classification of normal and kidney stone cases using a dataset consisting of 1799 CT images. Their approach achieved an impressive accuracy of 96.82 %, highlighting the efficacy of the XResNet-50. In (Jendeberg et al., 2021), they utilized a DL model trained on a dataset of 384 calcifications. Through a comparison with the performance of seven radiologists, they discovered that the proposed CNN model acheived an impressive accuracy rate of 93 %. In the study conducted by (Sudharson and Kokil, 2020), they utilized pre-trained networks to extract features from ultrasound images. These features were then classified using a SVM, and the final predictions were generated using the majority voting technique. Remarkably, the authors achieved an accuracy of 95.58 % in their classification task. Perrot et al. (De Perrot et al., 2019) conducted a study that employed radiomics feature extraction and the AdaBoost for kidney stone classification. Their study utilized a dataset consisting of 369 cases of kidney stones. Through their approach, they achieved 85.1 % results. The authors (Blau et al., 2018) developed a DL model specifically designed for the detection of cysts. Through their study, they achieved a true positive rate of 84.30 %. Alzu'bi et al. (Alzu'bi et al., 2022) conducted a study where they introduced a new dataset consisting of 8,400 images from 120 adult patients. The objective was to develop detection models using the ResNet50 architecture. Their study achieved impressive accuracy rate of 97 %. Islam et al. (Islam et al., 2022) implemented six models, three of which were based on Vision transformers, while the other three were based pre-trained architectures with adjustments in the last layers. Among these models, the swin transformer demonstrated superior performance, achieving an impressive accuracy of 99.30 %. Within the existing body of literature on kidney disease diagnosis, several notable challenges have been identified in prior methodologies. One significant limitation lies in the struggle to attain high accuracy rates, posing a challenge to the precision and reliability of disease detection mechanisms. Furthermore, the constraints associated with small datasets have been a recurring issue, impeding the ability of models to generalize effectively to diverse cases. Another noteworthy drawback is the prevalent reliance on singular model architectures, neglecting the potential advantages that can be derived from combining the strengths of multiple models to capture a more comprehensive spectrum of features. Moreover, an underexplored facet in the literature is the limited application of metaheuristic algorithms for optimizing ensemble models. These algorithms have the potential to enhance the synergistic interaction between models, improving overall performance. Recognizing these shortcomings, our research endeavors to address these challenges by proposing innovative solutions. Our approach involves leveraging ensemble techniques, specifically employing metaheuristic algorithms for optimal model combination, and augmenting datasets to mitigate the limitations associated with small data sizes. Through these strategies, we aim to significantly enhance the accuracy, reliability, and generalizability of kidney disease diagnosis models, contributing to the advancement of the field.

1.1. Novelty and research contributions

Kidney stone disease poses significant health risks, requiring accurate and timely diagnosis to save patients' lives. In this context, the use of DL models for kidney stone detection has gained attention. However, the limitations of single models, such as overfitting and low accuracy, have motivated the exploration of ensemble methods to overcome these challenges. This paper contributes to the field by proposing two novel ensemble models: StackedEnsembleNet and PSOWeightedAvgNet. These models address the limitations of single models and aim to enhance the accuracy and reliability of kidney stone classification. StackedEnsembleNet leverages the collective knowledge of four base models to capture complementary information and improve classification performance. On the other hand, PSOWeightedAvgNet utilizes PSO to optimize the weights of the weighted average ensemble, further enhancing the model's performance. The proposed models demonstrate superior performance in kidney stone detection compared to previous methods. By combining the predictions of multiple models or optimizing the weights, our models achieve higher accuracy rates, thus improving the timely and accurate diagnosis of kidney stone disease. Overall, this research contributes to the advancement of kidney stone detection by introducing novel ensemble models that address the limitations of single models and provide improved accuracy and reliability in the diagnosis of this dangerous condition. In summary, this research work makes the following key contributions:

- Proposal of StackedEnsembleNet, a two-level deep stack ensemble model for kidney stone detection. This model combines the predictions of multiple base models using a concatenation merge technique, resulting in a more robust and accurate classification of kidney stone patients.
- Development of PSOWeightedAvgNet, an ensemble model that employs PSO to find the optimal weights of each base model. By optimizing the weights, PSOWeightedAvgNet enhances the performance

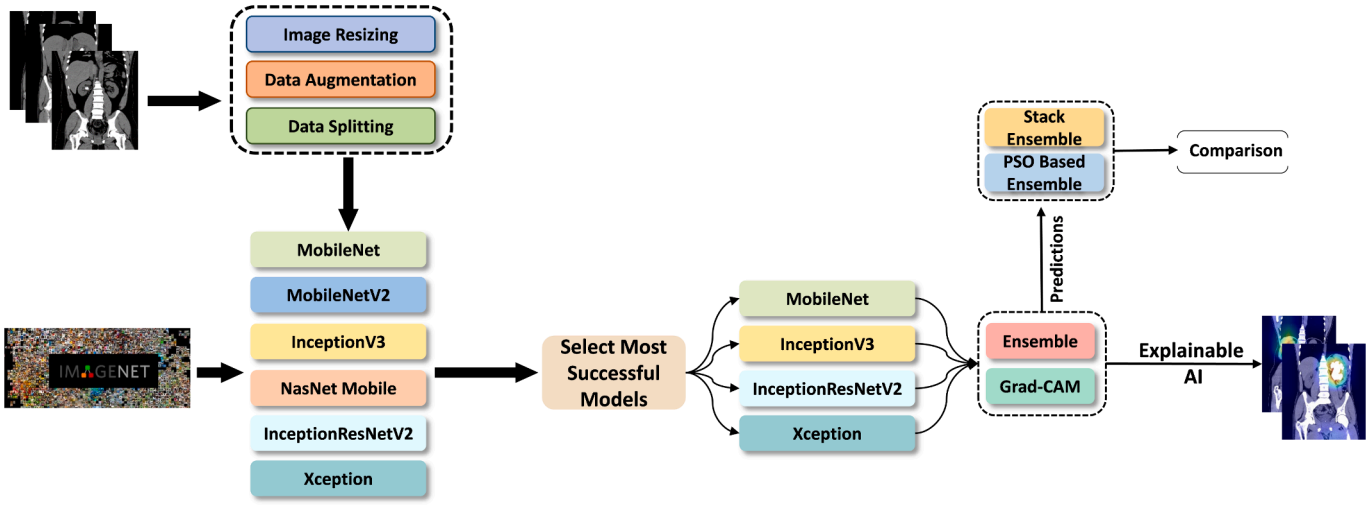

Fig. 1. Flowchart of the proposed methodology for classifying CT images.

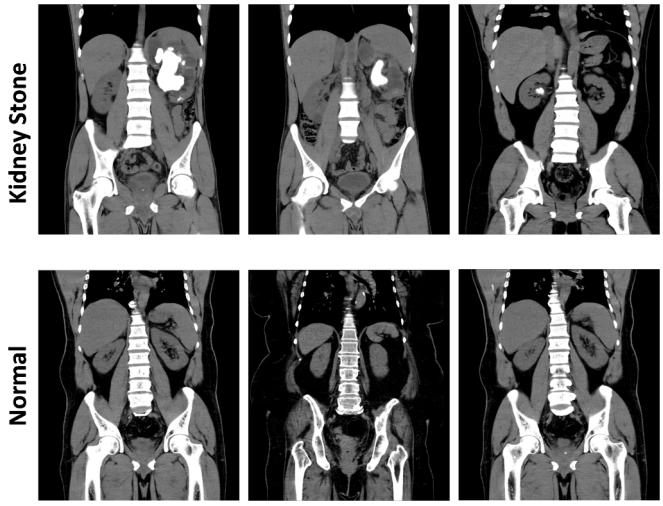

Fig. 2. Illustrative examples of CT images representing both individuals with kidney stones and those who are healthy.

and accuracy of the ensemble by determining the most effective combination of base model predictions for kidney stone classification.

- We perform comprehensive experiments on a sizable kidney CT dataset, incorporating the application of Grad-CAM for interpreting model decisions and offering visual explanations. Error case analysis is performed to gain insights into misclassifications. Additionally, the model's robustness is tested on an unseen dataset.
- We compare our proposed ensemble models, StackedEnsembleNet and PSOWeightedAvgNet, with previous methods in kidney stone detection. We demonstrate that our models outperform existing approaches, achieving higher accuracy and reliability in detecting kidney stones.

2. Materials and methods

In this section, we outline the methodology proposed for the classification of CT images. We introduce two ensemble models, namely StackedEnsembleNet and PSOWeightedAvgNet that were developed for improved classification performance. Fig. 1 presents the flowchart of the proposed methodology, offering a comprehensive overview of the entire process.

Table 1
Distribution of CT images and patients in each class.

Class	Patients	Images
Kidney Stone	278	790
Normal	165	1009

2.1. Dataset collection

For this study, a publicly available CT image dataset introduced by (Yildirim et al., 2021) was utilized. The dataset employed in this investigation comprises a total of 1799 CT images, divided into two distinct classes: 790 CT images describing patients with kidney stones, and 1009 CT images depicting individuals in good health. These images were gathered from a cohort of 433 individuals, encompassing 278 patients diagnosed with kidney stones and 165 subjects without any kidney abnormalities. These participants were admitted to Elazıg Fethi Sekin City Hospital in Turkey. The dataset serves as a valuable resource for training and evaluating models in the context of urinary stone disease. Fig. 2 showcases representative examples of CT images. Table 1 presents the distribution of images and patients across the two classes in the dataset. It provides a breakdown of the number of CT images available for kidney stones and healthy individuals, as well as the corresponding count of patients represented in each class.

2.2. Preprocessing and augmentation process

Preprocessing is a crucial step in DL, as it helps to optimize the input data for the model. In this study, all CT images were resized to a standardized dimension of 224x224 pixels. This resizing step ensures that the input images have consistent dimensions, which is essential for efficient model training and evaluation. For the dataset partitioning, an 80–20 split was applied, with 80 % of the images allocated for training and validation, and the remaining 20 % for testing the trained model's performance. This partitioning allows for robust model evaluation on unseen data and helps to estimate the model's generalization capability. In addition to preprocessing, image augmentation techniques were applied to enhance the training data and increase the model's ability to generalize to new samples. Image augmentation is particularly useful in medical imaging, as it helps to overcome the limited availability of annotated medical images. In this study, seven augmentation techniques were employed. By introducing diverse variations to the training images, these techniques enable the model to collect knowledge from a broader spectrum of data, thereby enhancing its proficiency in dealing

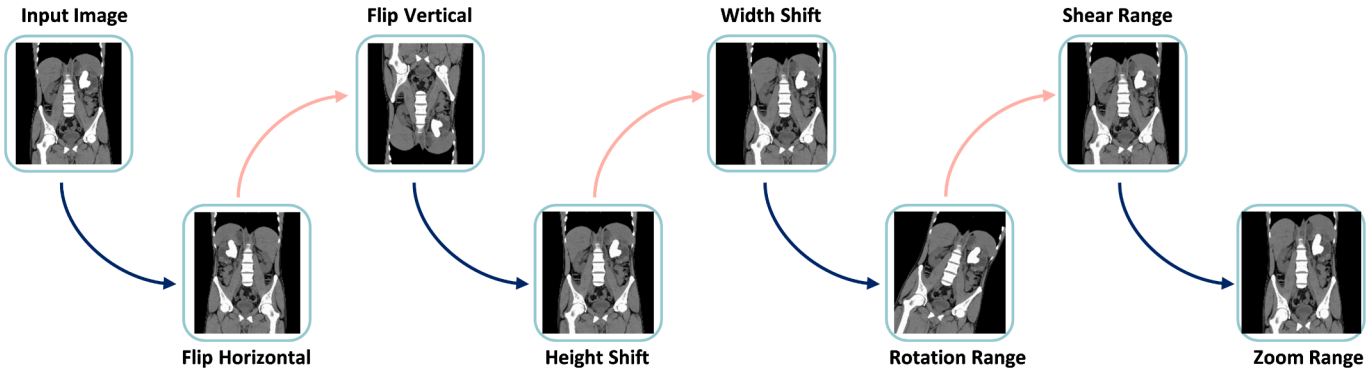

Fig. 3. Applied Augmentation Techniques on Input Image.

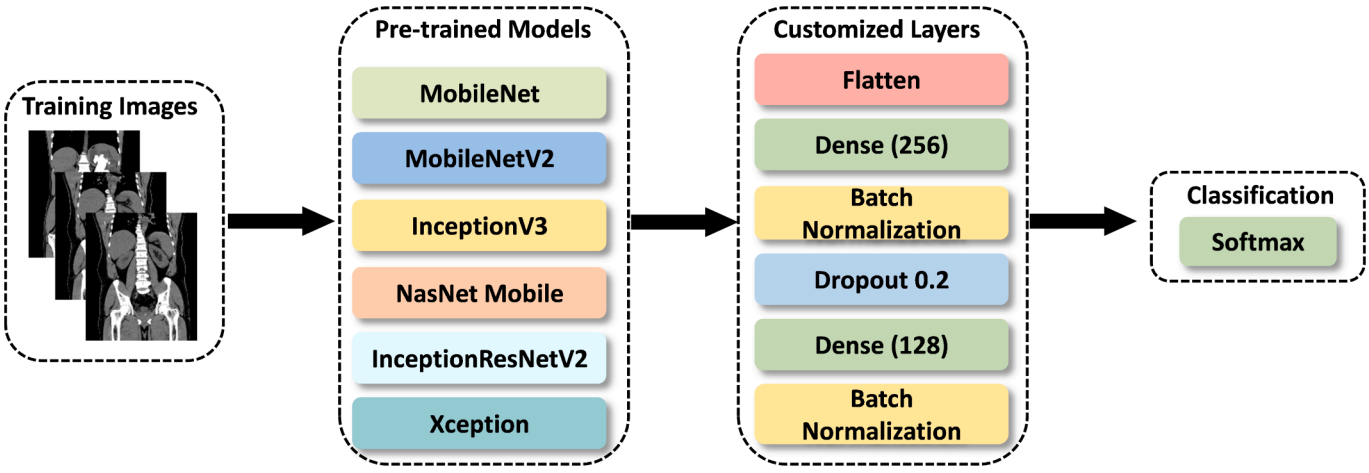

Fig. 4. Feature Extraction and Classification Process.

with variations and uncertainties encountered in real-world scenarios. Fig. 3 visually represents the applied augmentation techniques on an input image, providing a visual demonstration of the transformations performed during the augmentation process. Following the application of augmentation techniques, the quantity of training images depicting kidney stones expanded from 790 to 6320, and the number of normal images increased from 1009 to 8072.

2.3. *Feature extraction using transfer learning algorithms*

Feature extraction using pre-trained deep learning models is a valuable technique in medical image analysis. TL offers numerous advantages, notably the capability to harness knowledge acquired from large-scale datasets and apply it to specific tasks, even when the available data for those tasks is limited (Ahsan et al., 2023). In this study, four popular pre-trained models were utilized, and their architectures were modified to suit the kidney CT images dataset. As shown in Fig. 4, deep features were extracted from the pre-trained models. To extract these features, all layers except the final layer of the network were frozen. The output from the network was then flattened, resulting in a feature vector that captures the extracted features. To classify these features into their respective classes, a dense layer with 256 neurons was employed. Batch normalization (BN) was applied after this dense layer, which helps normalize the activations, leading to improved training stability and convergence. Moreover, a dropout rate of 0.2 was applied after the BN layer to reduce the risk of overfitting. This was achieved by randomly deactivating a fraction of neurons during the training process. Following the dropout layer, another dense layer with 128 neurons was employed. BN was applied once again to normalize the activations, ensuring a stable and consistent learning process. Finally, a dense layer was utilized for the classification, allowing the model to predict the presence of kidney stones or the absence of any abnormalities. The Table 2 provides an overview of the specifications for each model used in the study. The feature map dimensions, number of layers, and parameter counts are presented. A brief description of each model utilized in the study is presented below:

Table 2
Model Specifications.

Model	Feature Map	Layers	Parameters
InceptionV3	5, 5, 2048	48	34,944,930
InceptionResNetV2	5, 5, 1536	164	64,202,082
MobileNet	7, 7, 1024	28	16,108,866
Xception	7, 7, 2048	71	46,586,538

2.3.1. *InceptionV3*

InceptionV3 (Szegedy et al., 2016) is a deep architecture known for its efficient use of computational resources. The proposed architecture employs multiple convolutional layers with different filter sizes to capture diverse levels of features, a design that has found extensive application in various computer vision.

2.3.2. *InceptionResNetV2*

InceptionResNetV2 (Szegedy et al., 2017) is an enhanced version of InceptionV3 that incorporates residual connections. By combining the benefits of residual networks and the inception module, InceptionResNetV2 achieves improved performance in terms of accuracy and training

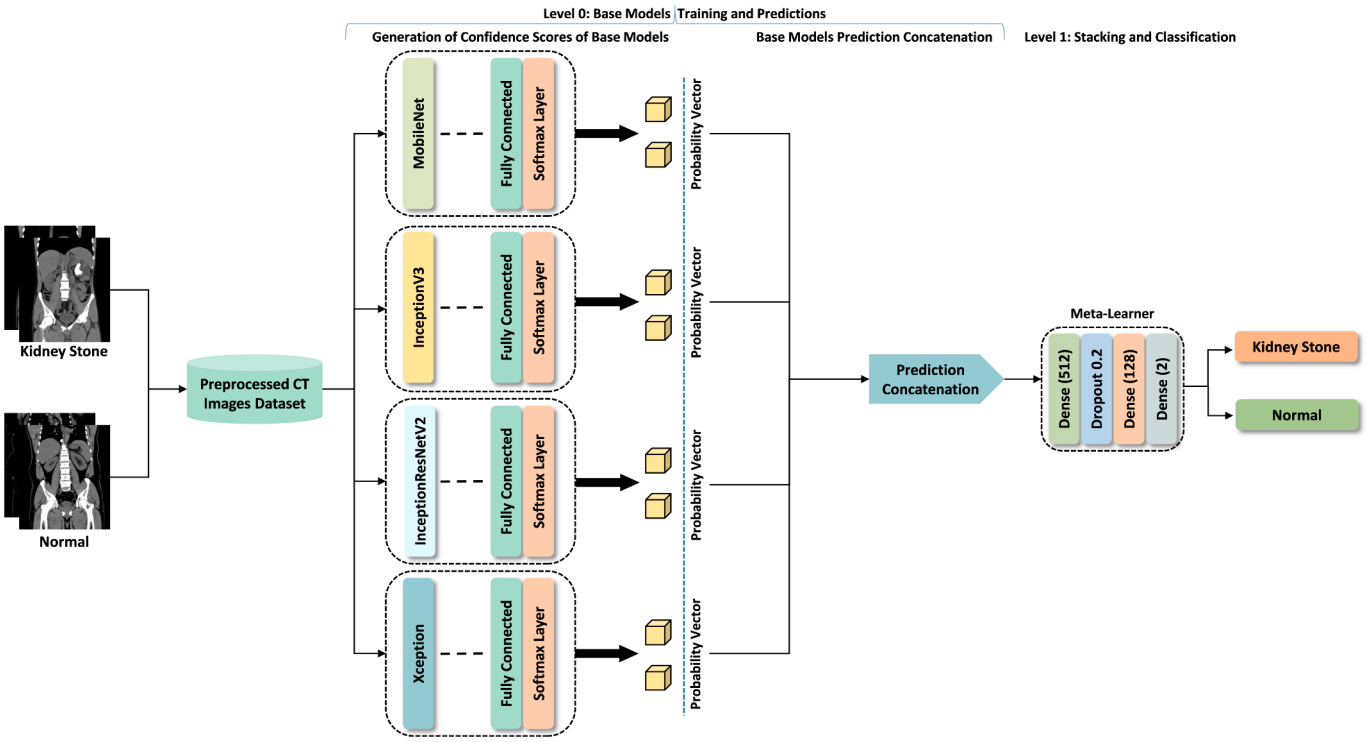

Fig. 5. StackedEnsembleNet architecture.

efficiency. It captures intricate features by using both parallel and residual connections, enabling better representation learning.

2.3.3. *MobileNet*

MobileNet (Howard et al., 1704) is a lightweight architecture designed specifically for mobile and resource-constrained devices. By incorporating depthwise separable convolutions, MobileNet achieves a commendable trade-off between model size and accuracy. This technique helps in reducing the number of parameters and computational complexity while retaining the ability to capture rich features.

2.3.4. *Xception*

Xception (Chollet, 2017) is a state-of-the-art architecture that emphasizes the depthwise separable convolutions. It replaces the traditional convolutional layers with depthwise separable convolutions, enabling the model to capture fine-grained spatial information while reducing the number of parameters. Xception has demonstrated strong performance in various image classification tasks, showcasing its effectiveness in learning rich representations from complex data.

2.4. *Proposed StackedEnsembleNet*

While single models have been commonly used in kidney stone classification, their limitations in achieving high accuracy and robustness have motivated the exploration of ensemble methods. In this section, we present the proposed StackedEnsembleNet, a two-level deep stack ensemble model, designed to overcome the shortcomings of single models and improve the classification of kidney stone patients. Ensemble learning techniques have gained popularity in various machine learning tasks by leveraging the collective knowledge of multiple models. Stack ensembles, in particular, have shown promising results due to their ability to capture complementary information from diverse base models (Polikar, 2012). The StackedEnsembleNet takes advantage of this concept to enhance the accuracy and reliability of kidney stone classification. The Fig. 5 illustrates the architecture of the StackedEnsembleNet, a two-level deep stack ensemble model for kidney stone classification.

At the first level of the StackedEnsembleNet, we train four TL models known for their effectiveness in extracting high-level features from complex images: InceptionV3, InceptionResNetV2, MobileNet, and Xception. Each base model is trained independently on the kidney CT images dataset, learning distinct representations of the data. To combine the predictions from these base models, we employ a concatenation merge technique.

2.4.1. *Concatenation of Feature Representations:*

The prediction vectors generated by each base model are concatenated to form a unified feature representation. This means that the outputs (feature vectors) from the final layers of each base model are joined end-to-end, resulting in a single, comprehensive feature vector. This unified representation integrates the diverse perspectives and knowledge captured by the individual models, allowing for a more robust ensemble model.

2.4.2. *Meta-Learner:*

Moving to the second level, known as the *meta*-learner, we utilize multiple layers for classification. The concatenated feature vectors obtained from the first level are fed into a dense layer with 512 neurons. This dense layer helps to further process and refine the combined features. Following this, a dropout rate of 0.2 is introduced to address overfitting concerns. Dropout regularization aids in generalization by randomly deactivating neurons during training, preventing the model from relying too heavily on specific features. After the dropout layer, another dense layer with 128 neurons is employed to further refine the feature representation. Finally, a dense layer with the appropriate number of output neurons is used for classification, allowing the StackedEnsembleNet to predict the presence or absence of kidney stones in the CT images.

2.4.3. *Aggregation of Predictions:*

By aggregating the predictions of these models through the concatenation and subsequent processing by the *meta*-learner, the

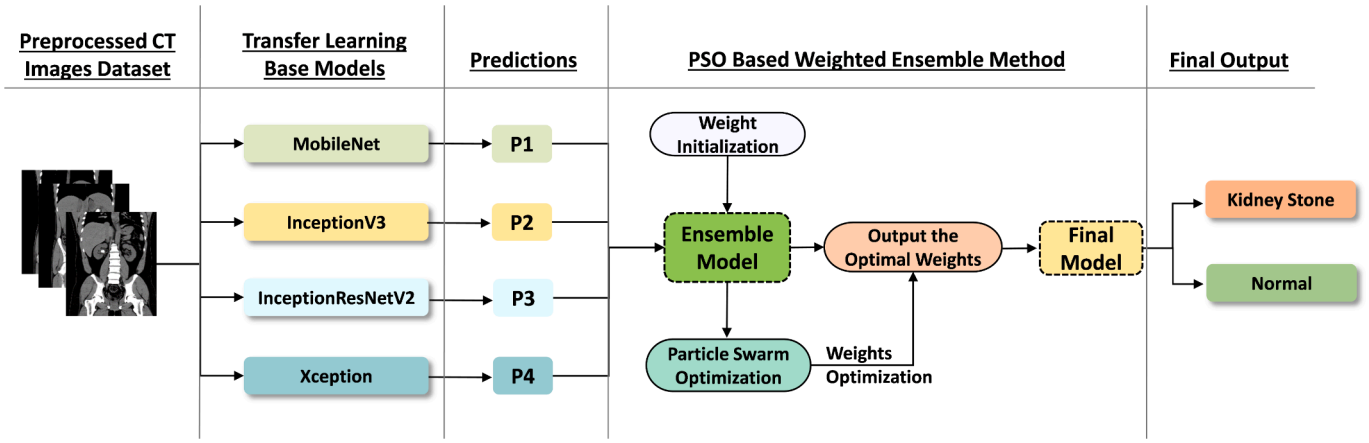

Fig. 6. The architecture of PSOWeightedAvgNet.

StackedEnsembleNet captures a more comprehensive understanding of the underlying patterns in kidney stone images. This approach enhances the model's accuracy and robustness compared to individual models. In conclusion, the proposed StackedEnsembleNet introduces a two-level deep stack ensemble model for kidney stone classification. Through the integration of knowledge from multiple base models and the aggregation of their predictions, the StackedEnsembleNet endeavors to improve the accuracy and reliability of kidney stone detection. The integration of diverse models and the utilization of advanced architectural components contribute to improved classification performance, paving the way for more effective diagnosis and treatment of kidney stone patients.

The proposed StackedEnsembleNet can be represented mathematically as follows:

Let:

- X Be the input kidney CT images dataset
- $M_1, M_2, M_3, and M_4$ be the four base models (InceptionV3, InceptionResNetV2, MobileNet, and Xception, respectively)
- $F_1, F_2, F_3, and F_4$ be the prediction vectors generated by the base models.
- F_{concat} be the concatenated feature representation obtained from the first level.
- D_1 be the dense layer with 512 neurons.
- DO be the dropout rate of 0.2.
- D_2 be the dense layer with 128 neurons.
- D_{out} be the dense layer with the appropriate number of output neurons for classification.

The StackedEnsembleNet can be defined as follows:

1. First Level

$$M_1 X = F_1 \tag{1}$$

$$M_2 X = F_2 \tag{2}$$

$$M_3 X = F_3 \tag{3}$$

$$M_4 X = F_4 \tag{4}$$

$$F_{concat} = Concatenate([F_1, F_2, F_3, F_4]) \tag{5}$$

2. Second level or Meta-Learner

$$D_1(F_{concat}) = D_1(F_{concat}) \tag{6}$$

$$DO(D_1(F_{concat})) = F_{dropout} \tag{7}$$

$$D_2(F_{dropout}) = D_2(F_{dropout}) \tag{8}$$

$$D_{out}(D_2(F_{dropout})) = predictions \tag{9}$$

The StackedEnsembleNet takes the kidney CT images dataset as input and trains four base models independently to obtain their prediction vectors. The prediction vectors are then concatenated to form a unified feature representation. The feature representation is passed through the second level of the model, which consists of a dense layer a dropout layer and another dense layer. Finally, the output is obtained by passing the result through a dense layer which produces the predictions for kidney stone presence or absence.

2.5. Proposed PSOWeightedAvgNet

The proposed PSOWeightedAvgNet is an ensemble architecture designed to address the challenges associated with training single models in medical imaging, such as limited data availability and suboptimal model performance (Jakubovitz et al., 2017). By harnessing the principles of ensemble learning, our primary goal is to attain precise detection and classification of individuals afflicted with kidney stones, surpassing the effectiveness of current methodologies. In pursuit of this objective, we integrate a sophisticated weighted average ensemble technique (Zhang et al., 2014), which functions as a potent fusion mechanism, elevating the overall performance substantially. This strategic integration ensures a distinctive approach, contributing novel insights to the existing body of knowledge in kidney stone detection. In the PSOWeightedAvgNet, we utilize a metaheuristic method called PSO (Kennedy and Eberhart, 1995) to optimize the weights used in the weighted average ensemble. PSO, an optimization algorithm inspired by the social behavior of bird flocking and fish schooling, imitates the movement of particles within a multidimensional search space to effectively find the optimal solution. In our case, PSO is used to determine the best weights for combining the predictions obtained from the four individual models (Nabavi-Kerizi et al., 2010).

PSO offers several advantages for weight optimization in ensemble models. Firstly, it is a global optimization algorithm, meaning it is capable of finding the best solution across the entire search space, rather than getting stuck in local optima. This is particularly beneficial in complex problems like medical image classification, where the search space can be vast and intricate. Secondly, PSO is a population-based algorithm, which allows it to explore multiple solutions simultaneously. This characteristic enhances its ability to discover diverse and complementary weighting configurations, resulting in a more effective ensemble model. By considering a range of weight combinations, PSO

facilitates the integration of different perspectives captured by the individual models, leading to improved accuracy and robustness. In the PSOWeightedAvgNet, the predictions obtained from the four individual models are combined using a weighted average ensemble. To obtain the final ensemble prediction, each model's output is multiplied by a specific weight, and these weighted predictions are then averaged. The weights used for the ensemble are optimized through the PSO algorithm, which searches for the combination of weights that maximizes the overall performance of the ensemble. By incorporating the PSO optimization process and the weighted average ensemble technique, the PSOWeightedAvgNet aims to leverage the collective knowledge of the individual models and improve the performance for kidney stone detection. The ensemble model's performance is enhanced by the optimized weights, which allow for an optimal fusion of the predictions from the individual models. This integration of diverse models and the utilization of PSO for weight optimization contribute to the improved classification performance and provide a promising approach for medical image analysis tasks. Fig. 6 iluustrates the architecture of PSOWeightedAvgNet. It illustrates the ensemble model's structure, incorporating the PSO algorithm to optimize the weights for combining predictions from individual models.

The proposed PSOWeightedAvgNet can be represented mathematically as follows:

1. Particle initialization

Initialize the particle positions X_i and velocities VC_i for each particle i in the swarm:

$$X_i(0) = Initialpositionofparticlei \tag{10}$$

$$VC_i(0) = Initialvelocityofparticlei \tag{11}$$

2. Update particle velocity and Position

Update the velocity VC_i and position X_i of each particle i at each iteration t using the following equations:

$$VC_i(t+1) = w\hat{A}\cdot VC_i(t) + c_1\hat{A}\cdot r_1\hat{A}\cdot(P_i - X_i(t)) + c_2\hat{A}\cdot r_2\hat{A}\cdot\left(P_{global,best} - X_i(t)\right) \tag{12}$$

$$X_i(t+1) = Clamp(X_i(t) + VC_i(t+1), \min weight, \max weight \tag{13}$$

Where

- $W = 0.5$ is the inertia weight.
- $c_1 = 1 and c_2 = 2$ are the acceleration coefficients for personal best and global best, respectively.
- $r_1 and r_2$ are random numbers between 0 and 1.
- P_i is the personal best position of particle i.
- $P_{global,best}$ is the global best position among all particles.

3. Update personal best and global best Positions

- Update the personal best position P_i for each particle i and the global best position $P_{global,best}$ based on the current objective function values:

$$P_i = UpdateBestPosition(X_i(t+1), P_i) \tag{14}$$

$$P_{global,best} = UpdateGlobalBestPosition(X(t+1), P_{global,best}) \tag{15}$$

4. PSO objective Function

The objective function calculates the sum of false positives (fp) and false negatives (fn) based on the weighted average predictions. By minimizing this objective function, the PSO algorithm aims to find the optimal weights for the ensemble model that result in accurate classification of kidney stone presence in the CT images.

$def weighted\_ensemble\_objective(weights, preds)$:

$y\_true = test\_batches.classes$

$y\_pred = np.average(preds, axis = 0, weights = weights)$

$conf\_matrix = confusion\_matrix(y\_true, np.argmax(y\_pred, axis = 1))$

$fp = conf\_matrix.sum(axis = 0) - np.diag(conf\_matrix)$

$fn = conf\_matrix.sum(axis = 1) - np.diag(conf\_matrix)$

$return np.sum(fp + fn)$

5. PSO optimization Loop

Perform the PSO optimization loop for a maximum of $T = 200$ iterations with a population size of $N = population\_size$:

- Initialize the personal best positions P_i and global best position $P_{global,best}$ for each particle and the swarm, respectively.
- Iterate from $t = 0$ to $t = 199$:
- Update the particle velocities and positions using the equations in step 2.
- Update the personal best positions P_i and global best position $P_{global,best}$ using the equations in step 3.

6. Obtain the optimal weights and objective function Value

- After the PSO optimization loop completes, obtain the optimal weights $opt\_weights$ and the objective function value opt_value based on the best positions found:

$$opt\_weights = P_{global,best} \tag{16}$$

$$opt\_value = weighted\_ensemble\_objective(opt\_weights, test\_set\_preds) \tag{17}$$

2.5.1. Objective function

The objective function serves as a crucial component in optimizing the weights for the ensemble model (Naser and Alavi, 2021). Moreover, it serves as a fundamental metric for evaluating the ensemble's performance by quantifying the classification accuracy in identifying kidney stone presence in CT images. The objective function specifically calculates the sum of false positives (fp) and false negatives (fn) based on the weighted average predictions. The primary objective of this objective function is to minimize the value it produces. This reduction is desirable as it signifies an improvement in the ensemble model's ability to accurately classify patients as either having or not having kidney stones. By utilizing the PSO algorithm to optimize the weights, the objective function aids in fine-tuning the ensemble's configuration, leading to enhanced accuracy and reliability in kidney stone classification. The optimization process seeks to strike a balance between minimizing false positives and false negatives, aiming to achieve the highest possible accuracy. By minimizing the objective function, the ensemble model can effectively leverage the collective knowledge of its constituent models to capture the underlying patterns in kidney stone images and make more accurate predictions. The incorporation of this objective function into the PSOWeightedAvgNet algorithm represents a significant step towards achieving optimal performance in kidney stone classification. It allows for the exploration and refinement of different weight combinations, ultimately guiding the ensemble model towards improved accuracy and robustness.

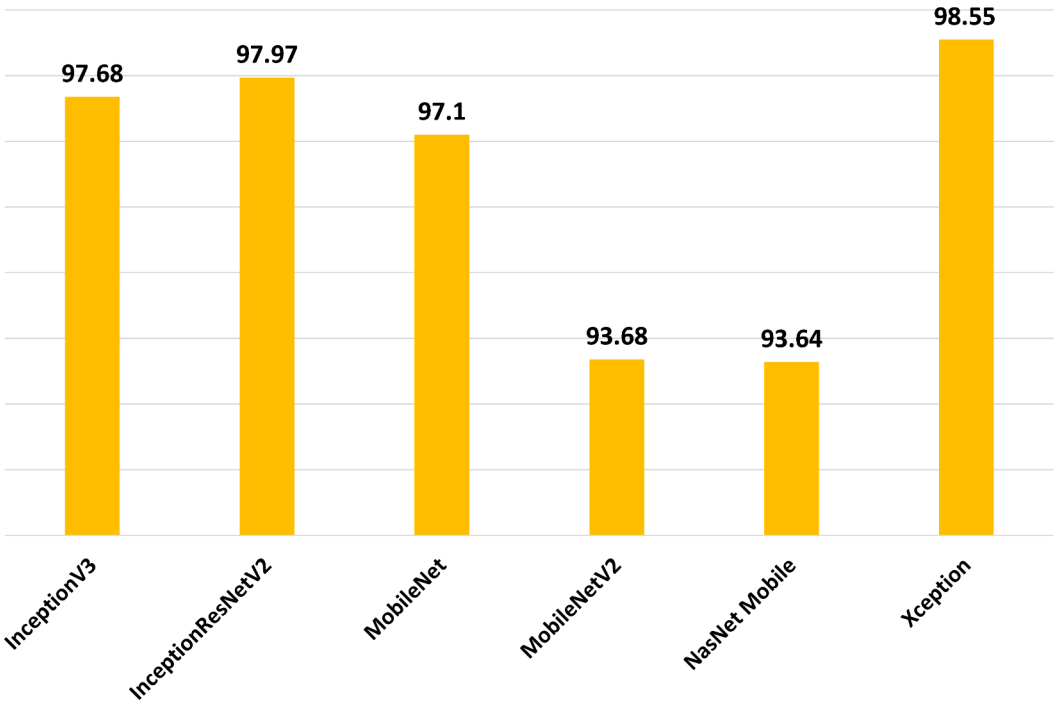

Fig. 7. Comparison of Accuracy among Six Individual Pre-Trained Models.

3. Experimental results and discussion

The experimental results and discussion section presents the evaluation metrics, hyperparameter analysis, and results analysis of the models. Interpretation of model decisions using Grad-CAM is discussed, along with error case analysis. The robustness of the model is assessed on an unseen dataset, and a comparison with previous methods is performed to highlight the model's superiority. The section concludes with a discussion on the implications of the findings and suggestions for future research directions.

3.1. Evaluation metrics

In the evaluation metrics section, multiple metrics are employed to assess the performance of the PSOWeightedAvgNet and StackedEnsembleNet model. Accuracy, Matthew's Correlation Coefficient (MCC), precision, recall, and F1-score are utilized to provide a comprehensive evaluation of the model's classification accuracy and predictive power. Additionally, the Receiver Operating Characteristic (ROC) curve is plotted to analyze the model's trade-off between true positive rate and false positive rate, further validating its performance. These metrics collectively offer insights into the effectiveness and reliability of the PSOWeightedAvgNet model in accurately classifying kidney stone patients.

$$Accuracy = \frac{tp + tn}{tn + fn + tp + fp} \tag{18}$$

$$Precision(PRC) = \frac{tp}{tp + fp} \tag{19}$$

$$Sensitivity(SENST) or Recall = \frac{tp}{tp + fn} \tag{20}$$

$$MCC = \frac{(tp \times tn - fp \times fn)}{\sqrt{(tp + fp)(tp + fn)(tn + fp)(tn + fn)}} \tag{21}$$

$$F1 - score(F1) = 2 \times \frac{PRC \times Recall}{PRC + Recall} \tag{22}$$

$$FalsePositiveRate(FPR) = \frac{FP}{FP + TN} \tag{23}$$

3.2. Hyperparameters

In the hyperparameters section, our experiments were conducted on Google Colab utilizing an Nvidia Tesla T4 GPU and 16 GB of RAM. Several crucial hyperparameters were carefully selected to ensure optimal performance of the models. During the training process, a learning rate of 0.001 was utilized in conjunction with the Adam optimizer, renowned for its efficiency in optimizing DL models. A batch size of 32 was chosen to balance computational efficiency and model convergence. The cross-entropy loss function was utilized, as it is commonly employed in multi-class classification tasks to measure the dissimilarity between predicted and actual class probabilities. In the StackedEnsembleNet architecture, the *meta*-learner is an essential component trained to combine the predictions of the base models. During training, a low learning rate of 0.0001 is employed specifically for the *meta*-learner. By using a lower learning rate, the *meta*-learner gradually updates its weights, allowing it to fine-tune the learned representations and adapt to the diverse perspectives captured by the individual base models. Regarding the PSO algorithm, key parameters were determined to facilitate the optimization of weights for the ensemble model. These included a population size of 25, acceleration coefficients of 1.0 (for personal best) and 2.0 (for global best), a maximum iteration count of 200, and an inertia weight of 0.5. These parameter values were selected based on empirical observations and prior studies to strike a balance between exploration and exploitation within the PSO optimization process, enhancing the ensemble model's accuracy and reliability in kidney stone classification.

3.3. Results analysis

In this section, we conduct an in-depth analysis of the results obtained from training various models on the dataset. The performance metrics of each individual model are presented, leading to the identification of the top-performing four models chosen for ensemble

Table 3
Performance Comparison of Individual Base Models with StackedEnsembleNet and PSOWeightedAvgNet.

Models	Accuracy	PRC	SENST	F1	MCC	FPR
InceptionV3	97.68 %	97.01 %	98.18 %	97.59 %	95.37 %	0.016
InceptionResNetV2	97.97 %	99.38 %	96.36 %	97.85 %	95.98 %	0.032
MobileNet	97.11 %	99.36 %	94.55 %	96.89 %	94.30 %	0.047
Xception	98.55 %	98.19 %	98.79 %	98.49 %	97.11 %	0.011
StackedEnsembleNet	**98.84 %**	**100 %**	**97.58 %**	**98.77 %**	**97.71 %**	**0.021**
PSOWeightedAvgNet	**98.84 %**	**98.79 %**	**98.79 %**	**98.79 %**	**97.68 %**	**0.011**

integration. Subsequently, we showcase the performance of the proposed ensemble model. Additionally, we conduct a comprehensive comparative analysis with pre-trained CNN backbones, providing insights into the relative efficacy and advantages of our ensemble approach.

3.3.1. *Models evaluation and selection for ensembling*

In this section, our exploration into the performance of six pre-trained models in kidney stone classification revealed distinctive accuracies, as illustrated in Fig. 7. Notably, InceptionV3 exhibited an accuracy of 97.68 %, showcasing its efficacy in capturing intricate features within kidney stone images. Similarly, InceptionResNetV2 achieved a commendable accuracy of 97.97 %, demonstrating its proficiency in discerning subtle patterns associated with kidney stones. MobileNet, with an accuracy of 97.1 %, affirmed its effectiveness in classification, benefitting from its lightweight architecture. However, the top performer was Xception, boasting an accuracy of 98.55 % and demonstrating exceptional capabilities in capturing nuanced details. Considering these individual strengths, InceptionV3, InceptionResNetV2, MobileNet, and Xception were selected for the ensemble models. Their collective knowledge and diverse architectural strengths made them well-suited for capturing a broad range of features and patterns in kidney stone images. This strategic ensemble approach aimed to enhance overall accuracy and robustness in kidney stone detection, capitalizing on the strengths of each selected model.

3.3.2. *Performance analysis of the proposed ensemble model*

After selecting the four best individual models, we designed two ensemble architectures to further improve the accuracy and robustness of kidney stone classification. The StackedEnsembleNet combines the predictions of the InceptionV3, InceptionResNetV2, MobileNet, and Xception models using a concatenation merge technique. This ensemble architecture aims to capture the diverse perspectives and feature representations of the individual models, resulting in improved

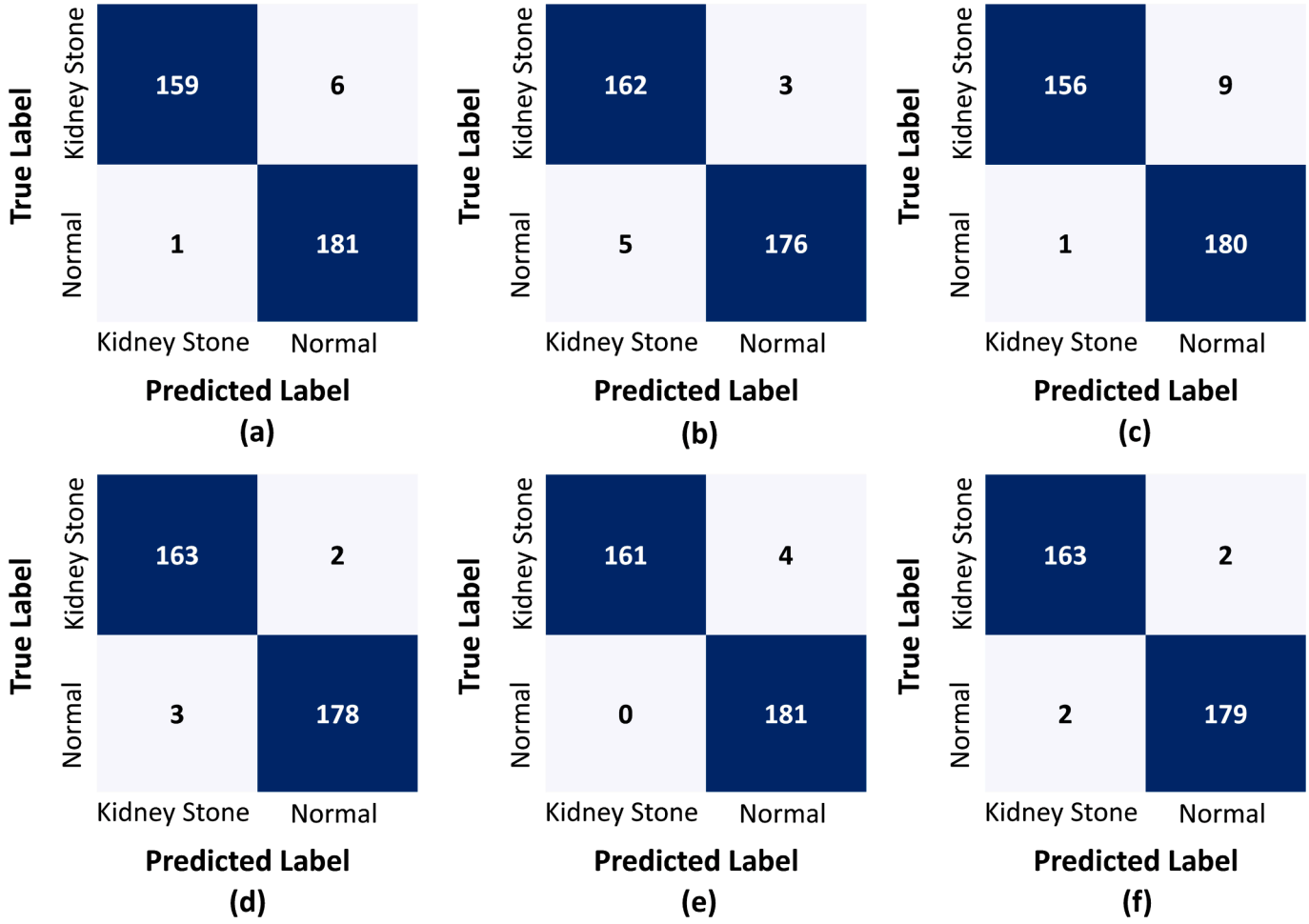

Fig. 8. Confusion Matrix Comparison of Four Base Models and Ensemble Models (a) InceptionResNetV2, (b) InceptionV3, (c) MobileNet, (d) Xception, (e) StackedEnsembleNet and (f) PSOWeightedAvgNet.

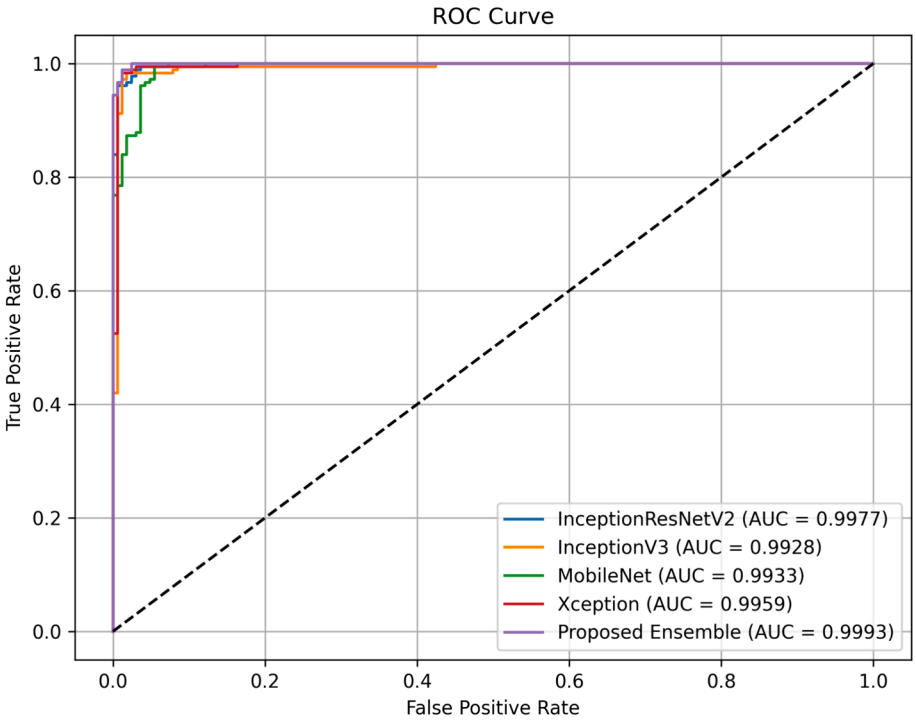

Fig. 9. ROC Curve Comparisons of PSOWeightedAvgNet with Base Models.

classification performance. The PSOWeightedAvgNet ensemble architecture utilizes the PSO algorithm to optimize the weights for combining the predictions of the individual models. By leveraging the collective knowledge of the InceptionV3, InceptionResNetV2, MobileNet, and Xception models through weighted averaging, the PSOWeightedAvgNet achieves enhanced accuracy and reliability in kidney stone classification. Table 3 presents a comparison of the performance of the individual base models with the StackedEnsembleNet and PSOWeightedAvgNet ensemble models. The ensemble models outperform the individual models across various performance metrics, including accuracy, precision, sensitivity, F1-score, and MCC. When comparing the two ensemble models, it is observed that the PSOWeightedAvgNet has higher sensitivity and F1-score compared to the StackedEnsembleNet. The PSOWeightedAvgNet achieves a sensitivity and F1-score of 98.79 %, indicating its ability to accurately detect positive cases of kidney stones and achieve a balance between precision and recall. On the other hand, the StackedEnsembleNet achieves an accuracy of 98.84 % and 100 % precision, showcasing its high overall accuracy and precise

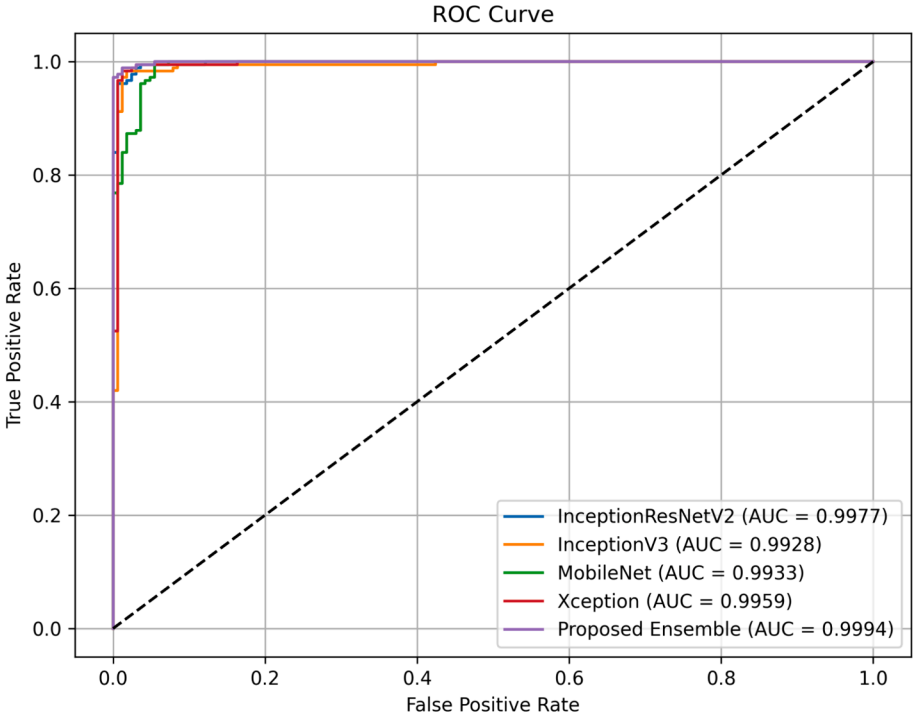

Fig. 10. ROC Curve Comparisons of StackedEnsembleNet with Base Models.

Table 4
Comparison of Performance Metrics: StackedEnsembleNet and PSOWeightedAvgNet against Selected CNN Backbones.

Models	Accuracy	PRC	SENST	F1	MCC
DenseNet169	96.55 %	93.98 %	98.73 %	96.30 %	93.17 %
DenseNet201	96.28 %	93.98 %	98.11 %	96.00 %	92.59 %
ResNet50V2	96.00 %	92.86 %	98.73 %	95.71 %	92.12 %
ResNet101V2	95.73 %	96.30 %	94.55 %	95.41 %	91.43 %
MobileNetV2	94.65 %	92.86 %	95.71 %	94.26 %	89.28 %
NasNet Mobile	93.85 %	91.76 %	95.12 %	93.41 %	87.71 %
VGG16	96.55 %	96.30 %	96.30 %	96.30 %	93.07 %
StackedEnsembleNet	**98.84 %**	**100 %**	**97.58 %**	**98.77 %**	**97.71 %**
PSOWeightedAvgNet	**98.84 %**	**98.79 %**	**98.79 %**	**98.79 %**	**97.68 %**

identification of kidney stone cases. Considering the higher sensitivity and F1-score, the PSOWeightedAvgNet can be regarded as the better-performing ensemble model for accurately detecting the presence of kidney stones. However, it is important to note that both ensemble models demonstrate excellent performance, highlighting the effectiveness of ensemble learning techniques in improving the classification of kidney stone patients.

The Fig. 8 presents the confusion matrices for the individual base models (InceptionV3, InceptionResNetV2, MobileNet, and Xception) as well as the StackedEnsembleNet and PSOWeightedAvgNet ensemble models. Each confusion matrix illustrates the classification results for a total of 346 test images. The performance of the individual base models is reflected in the number of misclassified images. InceptionV3 misclassifies 7 images, InceptionResNetV2 misclassifies 8 images, MobileNet misclassifies 10 images, and Xception misclassifies 5 images. Among the ensemble models, the StackedEnsembleNet misclassifies 4 images, with all 4 misclassifications occurring in the kidney stones class. On the other hand, the PSOWeightedAvgNet misclassifies 4 images, with 2 misclassifications in the kidney stone class and 2 misclassifications in the normal class.

Figs. 9 and 10 showcase the ROC curves that compare the performance of the ensemble models, PSOWeightedAvgNet and StackedEnsembleNet, with the individual base models. In Fig. 9, the PSOWeightedAvgNet ensemble model demonstrates exceptional performance, as evidenced by the high AUC values. InceptionResNetV2 achieves an AUC of 99.77 %, InceptionV3 achieves an AUC of 99.28 %, MobileNet achieves an AUC of 99.33 %, Xception achieves an AUC of 99.59 %, and the PSOWeightedAvgNet surpasses all of them with an impressive AUC of 99.93 %. Similarly, in Fig. 10, the StackedEnsembleNet ensemble model exhibits outstanding performance with higher AUC values compared to the individual base models. The StackedEnsembleNet outperforms all individual models with a remarkable AUC of 99.94 %. Both the PSOWeightedAvgNet and StackedEnsembleNet demonstrate superior performance, achieving higher AUC values and showcasing their effectiveness in accurately classifying kidney stones.

3.4. Comparison with pre-trained CNN backbones

The proposed ensemble models demonstrated commendable performance for kidney stone detection, as evaluated through various metrics. To further emphasize their efficacy, we conducted a comprehensive comparison with seven state-of-the-art CNN backbones. Table 4 provides a detailed overview of the accuracy and other performance metrics for each model. The StackedEnsembleNet and PSOWeightedAvgNet consistently outperform individual pre-trained CNN backbones across all metrics. Notably, the ensemble models achieve the performance, showcasing their robustness and superiority in kidney stone classification. The ensemble approach effectively leverages the diverse strengths of the constituent models, resulting in a more accurate and reliable diagnostic tool compared to standalone CNN backbones. Our research outcomes definitively confirm that purposeful implementation of ensembling techniques proficiently capitalizes on the synergies and complementary information inherent in a diverse range of CNN models.

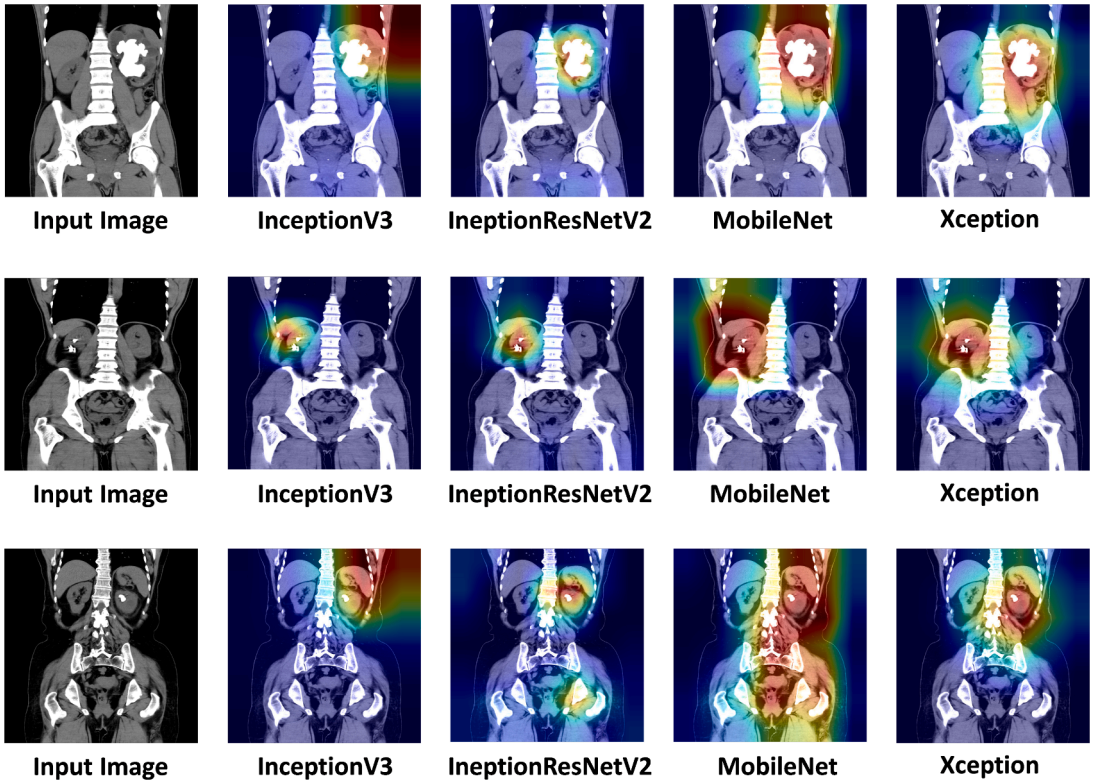

Fig. 11. Grad-CAM Visualization of Base Models on Kidney Stone Images.

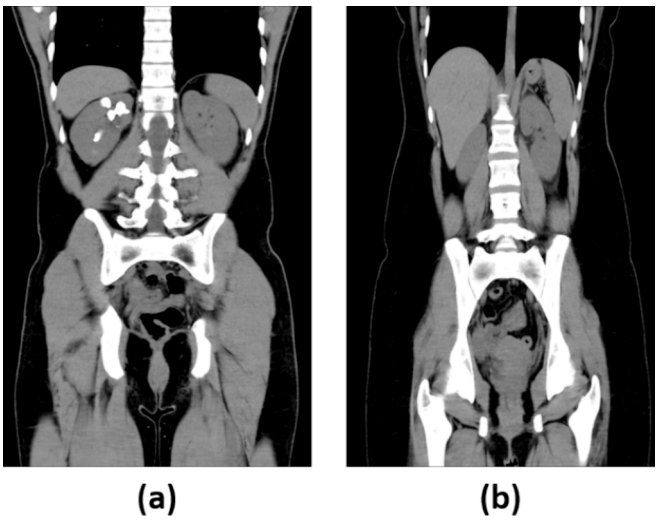

Fig. 12. Examples of Error Cases in Kidney Stone Classification.

3.5. Interpreting model decisions with Grad-CAM

The Grad-CAM technique is employed to interpret and visualize the decisions made by the models in the context of medical imaging. By utilizing Grad-CAM, we gain valuable insights into the areas of the image that have the most significant impact on the classification decision, effectively highlighting the regions the model focuses on during the classification process. This technique is particularly useful in the medical field as it aids in understanding the features and patterns that drive the model's decision-making process. In Fig. 11, we present the Grad-CAM visualizations of the four base models (InceptionV3, InceptionResNetV2, MobileNet, and Xception) on a selection of kidney stone images. By examining the Grad-CAM outputs, we gain valuable insights into the models' attention and focus on specific areas of the CT images. Notably, it is observed that all four models predominantly emphasize the regions corresponding to the presence of kidney stones in the CT scans. By showcasing the Grad-CAM visualizations of the base models that were utilized in constructing the ensemble models, we provide a deeper understanding of the underlying features and patterns that contribute to the models' decision-making process. This analysis aids in validating the models' ability to effectively detect and classify kidney stones based on the identified stone areas in the CT images.

3.6. Error case analysis

Despite achieving high performance accuracy in kidney stone detection, it is crucial to analyze the cases where our proposed ensemble models, StackedEnsembleNet and PSOWeightedAvgNet, fail to predict the correct class of CT images. While these models exhibit an impressive classification accuracy of 98.84 %, there are instances where misclassifications occur, posing potential risks due to the severe nature of the disease. In Fig. 12, we present examples of CT images where the ensemble models fail to classify the images correctly. In Fig. 12(a), kidney stone samples are depicted, where StackedEnsembleNet fails to detect the presence of kidney stones, while PSOWeightedAvgNet successfully classifies the images. Conversely, in Fig. 12(b), samples of normal patients are shown, where PSOWeightedAvgNet fails to detect the absence of kidney stones, while StackedEnsembleNet accurately classifies the images. Analyzing and understanding these error cases is crucial for further refinement and improvement of the ensemble models. By identifying the specific scenarios where misclassifications occur, further investigations can be conducted to enhance the models' performance and mitigate the risk of misdiagnosis in critical cases.

3.7. Robustness assessment on an unseen dataset

In this section, we conduct an evaluation of the PSOWeightedAvgNet and StackedEnsembleNet models on an unseen dataset without any additional training. This evaluation poses a significant challenge as it assesses the models' capacity to generalize to external data that was not part of the training process. To conduct this assessment, we collected 254 samples of CT images from the Kaggle repository "CT KIDNEY DATASET," sourced from various hospitals in Dhaka, Bangladesh (Islam et al., 2022). Table 5 presents the results of the four base models along with the PSOWeightedAvgNet and StackedEnsembleNet on this unseen dataset. The evaluation metrics include accuracy, precision (PRC), sensitivity (SENST), F1-score, and MCC. The results demonstrate the robustness of the ensemble models on the unseen dataset. InceptionV3 achieves an accuracy of 97.24 %, InceptionResNetV2 achieves an accuracy of 97.64 %, MobileNet achieves an accuracy of 95.67 %, and Xception achieves an accuracy of 97.64 %. Comparatively, the StackedEnsembleNet and PSOWeightedAvgNet models outperform the individual base models, both achieving an accuracy of 98.43 %. These ensemble models also exhibit higher precision, sensitivity, F1-score, and MCC values, indicating their effectiveness in accurately classifying kidney stones in the unseen dataset. This robustness assessment on an external dataset further validates the generalization capability of the ensemble models, showcasing their potential for reliable kidney stone detection in real-world scenarios.

3.8. Ablation study

In this section, we conducted an ablation study to systematically examine the impact of specific hyperparameters, including batch size and learning rate, on the performance of our models. To delve into this analysis, we tested four distinct models, varying the batch size between 8, 16, and 32, and exploring learning rates of 1e-3 and 1e-4. The results of this ablation study are illustrated in Fig. 13, where it becomes evident that the model with a batch size of 32 and a learning rate of 1e-3 consistently outperformed the other configurations. This observation underscores the significance of these specific hyperparameters in influencing the overall effectiveness and performance of the models, providing valuable insights for optimizing their configuration in practical applications.

In our experimentation, we evaluated the performance of four different optimizers, and the corresponding accuracy results for each model are depicted in Fig. 14. Notably, Adam optimizer consistently outperformed the other optimization algorithms, prompting its selection for our proposed model. The superiority of Adam optimizer can be

Table 5
Performance Comparison of Base Models, PSOWeightedAvgNet, and StackedEnsembleNet on Unseen Dataset.

Models	Accuracy	PRC	SENST	F1	MCC	FPR
InceptionV3	97.24 %	97.74 %	97.01 %	97.38 %	94.48 %	0.0331
InceptionResNetV2	97.64 %	99.23 %	96.27 %	97.73 %	95.32 %	0.0401
MobileNet	95.67 %	99.20 %	92.54 %	95.75 %	91.58 %	0.0772
Xception	97.64 %	98.48 %	97.01 %	97.74 %	95.28 %	0.0328
StackedEnsembleNet	**98.43 %**	**100 %**	**97.01 %**	**98.48 %**	**96.89 %**	**0.0323**
PSOWeightedAvgNet	**98.43 %**	**100 %**	**97.01 %**	**98.48 %**	**96.89 %**	**0.0323**

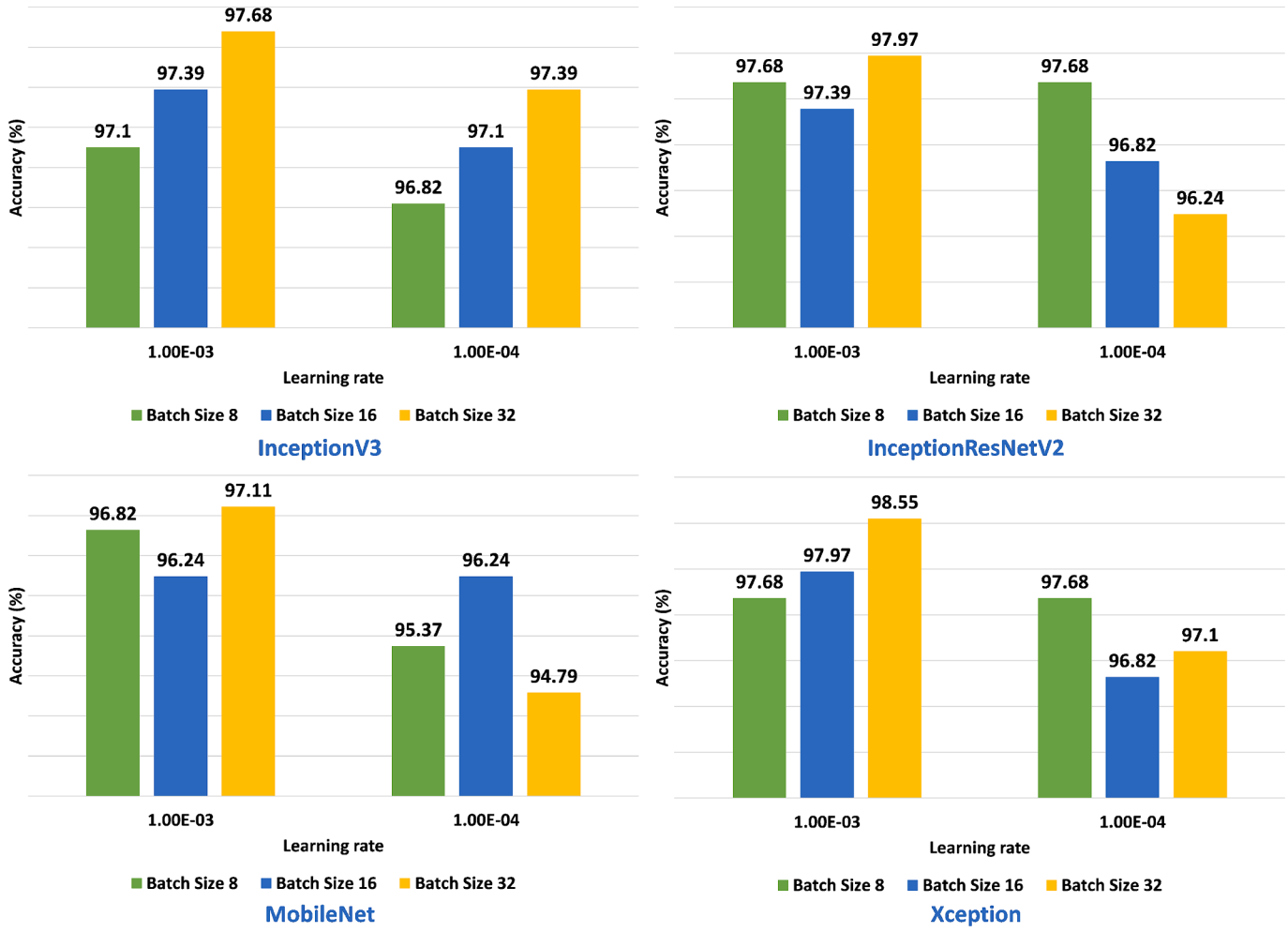

Fig. 13. Ablation Study Results.

attributed to its adaptive learning rate mechanism, which dynamically adjusts the learning rates for each parameter during training. This adaptability allows Adam to converge faster and navigate complex optimization landscapes effectively. Additionally, Adam combines the advantages of both momentum and RMSprop, offering a robust optimization approach. The empirical evidence from our study suggests that Adam optimizer aligns well with the characteristics of our dataset and task, contributing to improved accuracy and convergence during training.

3.9. Comparison with previous methods

To evaluate the efficacy of our proposed PSOWeightedAvgNet and StackedEnsembleNet models for kidney stone detection, we conducted a comparative analysis against previously reported methods in the literature. Table 6 provides a summary of the comparison, showcasing the accuracy achieved by each method. In terms of accuracy, our proposed models, StackedEnsembleNet, and PSOWeightedAvgNet, demonstrate superior performance compared to previous methods, achieving an impressive accuracy of 98.84 %. In contrast, the previous methods reported accuracy values ranging from 94.67 % to 97.50 %. The distinctive advantage of our proposed method stems from its ensemble framework, where the collective predictive capabilities of multiple base models are harnessed to yield superior performance. By leveraging the diverse perspectives captured by these models, the ensemble models can capture a more comprehensive understanding of the underlying patterns in kidney stone images, leading to improved accuracy and robustness. Furthermore, the utilization of advanced architectural components, such as *meta*-learning and PSO optimization, further enhances the performance of our models. The superior performance of our proposed models can be attributed to their ability to effectively integrate and aggregate the predictions from individual base models, leveraging their complementary knowledge and capturing a broader range of features relevant to kidney stone detection. This ensemble-based approach allows our models to make more accurate and reliable predictions, demonstrating their potential for advancing the field of kidney stone detection and contributing to more effective diagnosis and treatment strategies.

3.10. Discussion

The present study introduced two ensemble models, StackedEnsembleNet and PSOWeightedAvgNet, for kidney stone detection in CT images. Through extensive experimentation and evaluation, our proposed models demonstrated promising results, surpassing the performance of individual base models and achieving high accuracy and robustness in kidney stone classification. One key finding of our study is the effectiveness of ensemble learning in improving the accuracy and reliability of kidney stone detection. By combining the predictions of multiple base models, the ensemble models captured diverse perspectives and complementary information, leading to a more comprehensive understanding of the underlying patterns in kidney stone images. This integration of knowledge from different models helped to mitigate the limitations of individual models, resulting in improved performance. In

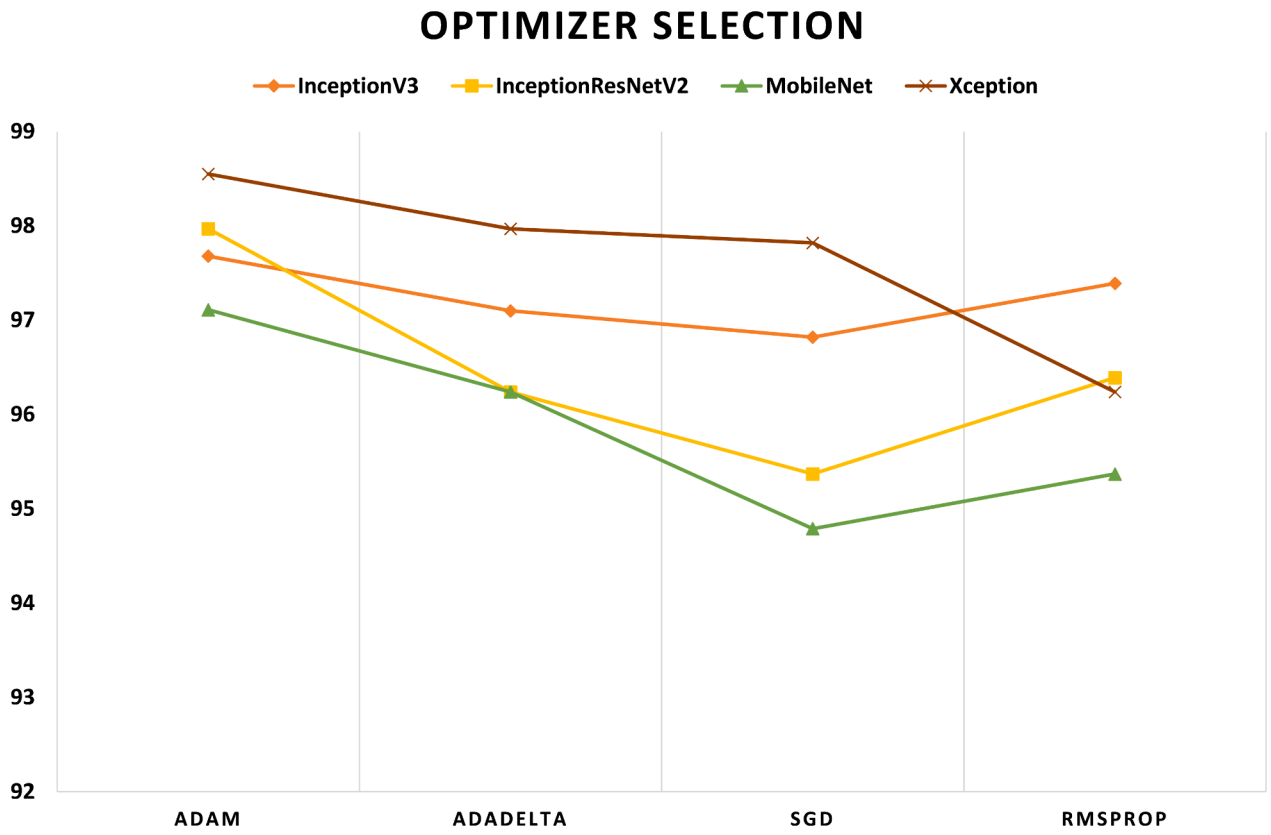

Fig. 14. Comparative Analysis of Different Optimizers.

Table 6
Performance Comparison of Proposed Models with Previous Methods for Kidney Stone Detection.

Reference	Method	Accuracy
(Yildirim et al., 2021)	XResNet-50	96.82 %
(Wu and Yi, 2020)	fusion-based model	94.67 %
(Sudharson and Kokil, 2020)	Hybrid model	95.58 %
(Alzu'bi et al., 2022)	ResNet50	97.00 %
(Rajinikanth et al., 2023)	Fused network	96.64 %
(Ma et al., 2020)	Modified ANN	97.50 %
(Asif et al., 2023)	StoneNet	97.98 %
Proposed Method	**StackedEnsembleNet**	**98.84 %**
	PSOWeightedAvgNet	**98.84 %**

terms of model performance, both StackedEnsembleNet and PSO-WeightedAvgNet exhibited impressive accuracy values, achieving 98.84 % accuracy on the test dataset. The ensemble models also demonstrated high precision, sensitivity, F1-score, and MCC values, indicating their ability to accurately classify kidney stone cases. Comparing our models with previous methods, our proposed ensemble models consistently outperformed existing approaches in terms of accuracy. This improvement can be attributed to the ensemble framework, which leverages the collective knowledge of multiple base models. The ensemble models excel in capturing diverse features and patterns, enabling them to make more accurate predictions compared to single models or fusion-based approaches. The utilization of advanced architectural components, such as *meta*-learning and PSO, further contributed to the performance enhancement of our models. The *meta*-learning framework in StackedEnsembleNet facilitated the learning of diverse representations from the base models, while the PSO optimization in PSOWeightedAvgNet effectively optimized the ensemble weights, resulting in more precise and robust predictions. We also conducted a robustness assessment on an unseen dataset to evaluate the generalization capability of our models. The results demonstrated the models' ability to perform well on external data without any additional training. This further strengthens the validity and potential applicability of our models in real-world scenarios. However, it is essential to recognize and address the limitations of our study. Despite achieving high accuracy, there were cases where the models failed to correctly classify the CT images, which can have critical implications in a medical setting. Further improvements and refinements are necessary to address these misclassifications and enhance the models' sensitivity and specificity. In conclusion, our proposed StackedEnsembleNet and PSO-WeightedAvgNet models offer a powerful solution for kidney stone detection. The ensemble approach, coupled with advanced architectural components, demonstrated improved accuracy, robustness, and generalization capability. These models have the potential to assist medical professionals in accurate diagnosis and treatment planning for kidney stone patients. Future research directions could focus on addressing the misclassification challenges and exploring additional ensemble techniques to further improve the models' performance and reliability.

3.11. Practical implementations of the proposed ensemble model

The proposed ensemble model for kidney stone detection holds significant potential for various practical applications in the medical field. Here are four practical scenarios where the model can be applied:

3.11.1. Clinical diagnosis Support

The ensemble model can serve as a valuable tool for clinicians in accurately diagnosing kidney stones from CT images. Its enhanced accuracy, as demonstrated in our experiments, can provide reliable support in identifying and classifying kidney stones, facilitating prompt and accurate diagnosis for medical practitioners.

3.11.2. Automated radiology Assistance

Incorporating the ensemble model into radiology workflows can streamline the process of analyzing CT scans. Automated kidney stone detection can assist radiologists by highlighting potential areas of interest, reducing the time and effort required for manual examination. This, in turn, enhances the efficiency of radiological assessments.

3.11.3. Telemedicine Applications

In telemedicine settings, where remote diagnosis and consultations are becoming increasingly prevalent, the proposed model can contribute to reliable preliminary assessments. Its ability to accurately detect kidney stones in CT images provides a valuable resource for remote healthcare providers, enabling timely and accurate diagnoses even in virtual healthcare environments.

4. Conclusion and Future work

The proposed work in this study focuses on addressing the challenges in accurate kidney stone detection. To enhance the accuracy and dependability of the classification process, we introduced two ensemble models, namely StackedEnsembleNet and PSOWeightedAvgNet. StackedEnsembleNet utilizes a two-level deep stack ensemble approach, combining the predictions of multiple base models to enhance overall performance. PSOWeightedAvgNet, on the other hand, leverages PSO to optimize the weights of the base model predictions, resulting in improved ensemble performance. The effectiveness of the proposed models was thoroughly evaluated through extensive experiments conducted on a sizable kidney CT dataset. The dataset consisted of 1799 CT images, providing a diverse range of cases for analysis. In addition to accuracy evaluation, Grad-CAM analysis was employed to interpret the model decisions, providing visual explanations and insights into the regions of focus for kidney stone detection. Error case analysis further deepened the understanding of misclassifications, aiding in potential improvements and error mitigation. StackedEnsembleNet achieved an accuracy of 98.84 %, while PSOWeightedAvgNet achieved the same accuracy level. These results highlight the effectiveness of the ensemble models in accurately identifying kidney stones in CT images. The models also exhibited higher precision, sensitivity, F1-score, and MCC, further emphasizing their superiority in classification performance. The findings of this research contribute to the field of kidney stone detection by providing innovative ensemble models that outperform existing approaches. The proposed models hold the potential to enhance the efficiency and reliability of diagnoses, ultimately leading to prompt and precise treatment decisions for patients. The integration of Grad-CAM analysis and error case analysis enhances the interpretability of the models and aids in identifying areas of improvement. Future work in this area can explore the generalizability of the models on diverse datasets and further refine the diagnostic capabilities for kidney stone detection. Additionally, these ensemble models have the potential to be applied to a wide range of medical imaging tasks and various diseases, presenting new prospects for expanded applications in the healthcare domain.

CRediT authorship contribution statement

Sohaib Asif: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Formal analysis, Data curation, Conceptualization. **Xiaolong Zheng:** Writing – review & editing, Visualization, Supervision, Investigation, Formal analysis. **Yusen Zhu:** Validation, Software, Formal analysis.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Ahsan, M.M., Uddin, M.R., Ali, M.S., Islam, M.K., Farjana, M., Sakib, A.N., Al Momin, K., Luna, S.A., 2023. Deep transfer learning approaches for Monkeypox disease diagnosis. Expert Syst. Appl., 119483

Alelign, T., Petros, B., 2018. Kidney stone disease: an update on current concepts. Advances in Urology 2018.

Alzu'bi, D., Abdullah, M., Hmeidi, I., AlAzab, R., Gharaibeh, M., El-Heis, M., Almotairi, K.H., Forestiero, A., Hussein, A.M., Abualigah, L., 2022. Kidney tumor detection and classification based on deep learning approaches: A new dataset in CT scans. Journal of Healthcare Engineering 2022.

Asif, S., Wenhui, Y., Jinhai, S., Ain, Q.U., Yueyang, Y., Jin, H., 2022. Modeling a fine-tuned deep convolutional neural network for diagnosis of kidney diseases from CT images. In: 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE, pp. 2571–2576.

Asif, S., Zhao, M., Tang, F., Zhu, Y., Zhao, B., 2023. Metaheuristics optimization-based ensemble of deep neural networks for Mpox disease detection. Neural Netw.

Asif, S., Zhao, M., Chen, X., Zhu, Y., 2023. StoneNet: An efficient lightweight model based on depthwise separable convolutions for kidney stone detection from CT images. Interdiscip Sci.

Asif, S., Awais, M., Khan, S.U.R., 2023. IR-CNN: Inception residual network for detecting kidney abnormalities from CT images. Network Modeling Analysis in Health Informatics and Bioinformatics 12, 35.

Baygin, M., Yaman, O., Barua, P.D., Dogan, S., Tuncer, T., Acharya, U.R., 2022. Exemplar Darknet19 feature generation technique for automated kidney stone detection with coronal CT images. Artif. Intell. Med. 127, 102274.

Blau, N., Klang, E., Kiryati, N., Amitai, M., Portnoy, O., Mayer, A., 2018. Fully automatic detection of renal cysts in abdominal CT scans. Int. J. Comput. Assist. Radiol. Surg. 13, 957–966.

Brisbane, W., Bailey, M.R., Sorensen, M.D., 2016. An overview of kidney stone imaging techniques. Nat. Rev. Urol. 13, 654–662.

Chollet, F., 2017. Xception: Deep learning with depthwise separable convolutions. Proc. IEEE Conf. Comput. Vis. Pattern Recognit. 1251–1258.

De Perrot, T., Hofmeister, J., Burgermeister, S., Martin, S.P., Feutry, G., Klein, J., Montet, X., 2019. Differentiating kidney stones from phleboliths in unenhanced low-dose computed tomography using radiomics and machine learning. Eur. Radiol. 29, 4776–4782.

Edvardsson, V.O., Indridason, O.S., Haraldsson, G., Kjartansson, O., Palsson, R., 2013. Temporal trends in the incidence of kidney stone disease. Kidney Int. 83, 146–152.

Gunasekara, T., De Silva, P.M.C., Ekanayake, E., Thakshila, W., Pinipa, R., Sandamini, P., Gunarathna, S., Chandana, E., Jayasinghe, S., Herath, C., 2022. Urinary biomarkers indicate pediatric renal injury among rural farming communities in Sri Lanka. Sci. Rep. 12, 8040.

A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, arXiv preprint arXiv:1704.04861, (2017).

Islam, M.N., Hasan, M., Hossain, M., Alam, M., Rabiul, G., Uddin, M.Z., Soylu, A., 2022. Vision transformer and explainable transfer learning models for auto detection of kidney cyst, stone and tumor from CT-radiography. Sci. Rep. 12, 1–14.

Jakubovitz, D., Giryes, R., Rodrigues, M.R., 2017. Generalization error in deep learning. In: Compressed Sensing and its Applications: Third International MATHEON Conference. Springer, pp. 153–193.

Jendeberg, J., Thunberg, P., Lidén, M., 2021. Differentiation of distal ureteral stones and pelvic phleboliths using a convolutional neural network. Urolithiasis 49, 41–49.

Kennedy, J., Eberhart, R., 1995. Particle swarm optimization. In: Proceedings of ICNN'95-international conference on neural networks. IEEE, pp. 1942–1948.

Ma, F., Sun, T., Liu, L., Jing, H., 2020. Detection and diagnosis of chronic kidney disease using deep learning-based heterogeneous modified artificial neural network. Futur. Gener. Comput. Syst. 111, 17–26.

Nabavi-Kerizi, S., Abadi, M., Kabir, E., 2010. A PSO-based weighting method for linear combination of neural networks. Comput. Electr. Eng. 36, 886–894.

Naser, M., Alavi, A.H., 2021. Error metrics and performance fitness indicators for artificial intelligence and machine learning in engineering and sciences, Architecture. Structures and Construction 1–19.

Parakh, A., Lee, H., Lee, J.H., Eisner, B.H., Sahani, D.V., Do, S., 2019. Urinary stone detection on CT images using deep convolutional neural networks: evaluation of model performance and generalization. Radiology. Artificial Intelligence 1.

Polat, H., Danaei Mehr, H., Cetin, A., 2017. Diagnosis of chronic kidney disease based on support vector machine by feature selection methods. J. Med. Syst. 41, 1–11.

Polikar, R., 2012. Ensemble learning. Ensemble machine learning: Methods and applications 1–34.

Rajinikanth, V., Vincent, P.D.R., Srinivasan, K., Prabhu, G.A., Chang, C.-Y., 2023. A framework to distinguish healthy/cancer renal CT images using the fused deep features. Front. Public Health 11.

Shen, D., Wu, G., Suk, H.-I., 2017. Deep learning in medical image analysis. Annu. Rev. Biomed. Eng. 19, 221–248.

Shlipak, M.G., Fried, L.F., Cushman, M., Manolio, T.A., Peterson, D., Stehman-Breen, C., Bleyer, A., Newman, A., Siscovick, D., Psaty, B., 2005. Cardiovascular mortality risk in chronic kidney disease: comparison of traditional and novel risk factors. JAMA 293, 1737–1745.

Sorokin, I., Mamoulakis, C., Miyazawa, K., Rodgers, A., Talati, J., Lotan, Y., 2017. Epidemiology of stone disease across the world. World J. Urol. 35, 1301–1320.

Sudharson, S., Kokil, P., 2020. An ensemble of deep neural networks for kidney ultrasound image classification. Comput. Methods Programs Biomed. 197, 105709.

Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., Bray, F., 2021. Global cancer statistics 2020: GLOBOCAN estimates of incidence and

mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. 71, 209–249.
Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826.
Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A., 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. Thirty-first AAAI conference on artificial intelligence.
Vupputuri, S., Soucie, J.M., McClellan, W., Sandler, D.P., 2004. History of kidney stones as a possible risk factor for chronic kidney disease. Ann. Epidemiol. 14, 222–228.
Wu, Y., Yi, Z., 2020. Automated detection of kidney abnormalities using multi-feature fusion convolutional neural networks. Knowl.-Based Syst. 200, 105873.
Yang, Y., Lv, H., Chen, N., 2022. A survey on ensemble learning under the era of deep learning. Artif. Intell. Rev. 1–45.
Yildirim, K., Bozdag, P.G., Talo, M., Yildirim, O., Karabatak, M., Acharya, U.R., 2021. Deep learning model for automated kidney stone detection using coronal CT images. Comput. Biol. Med. 135, 104569.
Zhang, Y., Zhang, H., Cai, J., Yang, B., 2014. A weighted voting classifier based on differential evolution. Abstract and applied analysis, Hindawi.